ENGLISHES IN A GLOBALIZED WORLD: EXPLORING CONTACT EFFECTS ON OTHER LANGUAGES

EDITED BY: Alexander Onysko and Peter Siemund

frontiers Research Topics

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ENGLISHES IN A GLOBALIZED WORLD: EXPLORING CONTACT EFFECTS ON OTHER LANGUAGES

Topic Editors:
Alexander Onysko, University of Klagenfurt, Austria
Peter Siemund, University of Hamburg, Germany

Table of Contents

Editorial: Englishes in a globalized world: Exploring contact effects on other languages

Alexander Onysko[1]* and Peter Siemund[2]

[1]Department of English, University of Klagenfurt, Klagenfurt, Austria, [2]Institute of English and American Studies, University of Hamburg, Hamburg, Germany

Editorial on the Research Topic
Englishes in a globalized world: Exploring contact effects on other languages

The diversity of Englishes around the world continues to stir the interest of researchers working in different fields of linguistics, language and communication. English in its myriad forms, uses and functions globally can be approached from multiple perspectives, such as linguistic descriptions on all levels of language, language policy, socio-critical analysis, language history, language teaching and many more. Among the many ways of investigating Englishes, the current Research Topic takes the perspectives of language contact and multilingualism to highlight the interaction of Englishes with other languages. Analyzing Englishes in their multilingual contexts has gained importance in research on world Englishes (see, e.g., Siemund and Leimgruber, 2021; Siemund, 2022), and the application of language contact in relation to multilingual realities helps to understand the multiple forms, uses and functions of Englishes.

In light of this, the sixteen contributions to the Research Topic cover a great range of diverse contexts in which English interacts with other languages and in which Englishes have been and continue to be shaped by contact with other languages. In general, the articles in this Research Topic can be situated in a language contact model of world Englishes (Onysko, 2016), which functions as an umbrella for the different aspects and contexts dealt with in this Research Topic. The model, Language Contact Typology of Englishes, describes the general contact settings that underly the diversity of Englishes across the world. The Language Contact Typology (LCT) postulates five different macro-types of Englishes that emerge from basic scenarios of language contact: (1) Global Englishes (GEs), which captures the status of English as the main voice of globalizing trends and developments in business, technology, communication and various cultural domains (e.g., music and film). In this function, Englishes, mostly from Western L1 settings, serve as main donor codes of English linguistic material, typically borrowings, that become integrated in major language communities. (2) Englishes in Multilingual

Constellations (EMCs), which refers to situations in which English is part of speakers' bi- and multilingual repertoires and usually plays an active role in people's everyday lives among one or more other languages. Postcolonial contexts where English coexists with other languages and multilingualism in situations of migration to mainly English-speaking territories and communities are two major examples of EMCs. (3) Learner Englishes (LEs) most prototypically arise in acquisitional settings, in which English is learned in institutional contexts while the language is not used as an everyday code among others in a speaker community. The importance of English as an international medium of communication has given the language prominence as a learner language worldwide and this status also boosts the role of Global Englishes as knowledge of the language can facilitate the borrowing of English elements in other languages. At the same time, the use of English as a lingua franca among interlocutors not sharing the same L1 background or other codes is to various extents influenced by their learner Englishes implemented in international exchange and often influenced by transfer features from the speakers' L1 and other language backgrounds. (4) Koiné Englishes (KEs) describes varieties that are shaped by dialect contact. This holds for both standardized and non-standardized L1 Englishes (e.g., American English, British dialects, New Zealand English, and so on) used in largely monolingual settings. L1 Englishes in contact with other languages can take in some, mostly lexical, borrowings from these languages (e.g., Māori loanwords in New Zealand English). (5) English-based Pidgins and Creoles (EPCs) refers to the particular contact scenario of Englishes having been formed due to restricted communicative needs of dominant English speakers and speakers of other languages, mostly in situations of trade or slavery. From a current point of view, EPCs have developed into fully functional codes that frequently remain in close contact with more standardized forms of English, leading to a process of variable decreolization, as in the contexts of Jamaica, the US (Gullah), the Bahamas, and Hawai'i, to name just a few.

When applying the LCT to conventionally labeled Englishes, two major trajectories of variation have to be highlighted. Firstly, as the LCT takes contact settings as a point of departure, forms and uses of English can both be analyzed on the individual speaker level as well as among a speaker community if the conditions of contact are shared. This means that actual manifestations of Englishes can be influenced by several contact scenarios at a time, and conventionally named Englishes can actually be shaped by a combination of different contact settings. For instance, Indian English, Nigerian English, Singapore English and many other X-Englishes, can be subject to variable contact settings, combining the types of multilingual, learner and global Englishes depending on the individual or community level of English use and exposure to the language (see Onysko and Siemund on English in multilingual contexts). Secondly, intersections between the major contact types are

also possible on an historical dimension. Irish English (see the contributions to the Research Topic by Kirk and Shimada), for example, was shaped in a bilingual contact setting of speakers shifting from Irish towards English, i.e., a context of EMC. From a current perspective, however, the contact features of Irish English that have been carried over from Irish have become conventionalized forms of Irish English and the language has become part of dialect contact scenarios (Koiné Englishes) while Irish English as an EMC remains possible in active bilingual communities of Gaelic and English speakers.

Considering the flexible, contact-centric approach to the diversity of Englishes across the world, the contributions to this Research Topic speak to several of the main contact types.

The majority of the articles exemplify the type of Global Englishes (GEs), in which English elements become incorporated in different recipient language communities. Kapo's research report, for example, provides an overview of the types of English loans and their integration into Albanian, following a descriptive tradition of anglicism research.

Hunt's study investigates the use of English loanwords in a German newspaper that is published for a bilingual audience of English and German as heritage language speakers in Australia. Among his findings, Hunt determines that anglicisms in texts taken from the German press agency (DPA) are more frequently flagged compared to articles published by local Australian journalists. Increased flagging can create an alienating effect in a fully bilingual, local readership in Australia.

The contribution by Scherling et al. takes the perception and reinterpretation of English song lyrics by native speakers of Japanese as a case in point to show how English elements are integrated into recipient language phonetic systems. The processes at play comprise sound substitutions, insertions, deletions and boundary transgressions with substitutions occurring by far most frequently in the author's corpus of *soramimi* ("mishearings").

Schuring and Zenner take a socio-pragmatic developmental approach to anglicism research. They investigate the use of English loanwords in Belgian Dutch among a group of young children using sociolinguistic interviews. Results show an overall rate of 9.7% of English loanwords with female children using significantly more English elements.

Research on aspects of the contact type of Global Englishes continues with two in-depth studies that each trace the integration and development of one English expression in a recipient language. Kailich follows the occurrence of the English-induced neologism *Covidiota* in Spanish Twitter posts, and De Pascale et al. take construction grammar as a lens to observe the life-cycle of the phrasal borrowing *pimp my ride* in Belgian Dutch. The section on Global Englishes closes with a perspective article by Schaefer who takes the example of anglicisms in the German radio to call for future anglicism research to focus on semiotic assemblages in transmedial and transmodal mass media communication.

Three studies in the Research Topic deal with aspects of Englishes in multilingual constellations (EMCs). McLellan investigates multilingual language practices in Malaysia and Brunei. He shows the interplay of English and Malay in examples of language use on social media, posing the question of whether social media are drivers of language change and whether the multilingual practices observable there might even be considered separate varieties.

Chan tackles the issue of constructional borrowing among bilingual English and Cantonese speakers in Hong Kong. The author provides a close up on three English constructions, which can occur in Hong Kong Cantonese when it functions as a matrix language in this multilingual constellation.

Westphal provides an analysis of multilingual contact effects in world Englishes on the level of pragmatics. He focuses on question tags in Nigerian English, taking both a corpus linguistic and a questionnaire-based survey approach. The findings of the corpus data demonstrate the use of invariant tags of both English and Nigerian origin while prescriptive attitudes emerge from the favorable ratings of variant tag questions.

A few more papers in this Research Topic are couched in multilingual settings that intersect with Learner Englishes and the use of English as a lingua franca. In line with transnationalism and the sociolinguistics of globalization (Blommaert, 2010), Mohr's study is an example of world Englishes research from a transnational and social-mediatized perspective. She investigates the use of English as a lingua franca in hashtags related to a popular tourist destination, tapping into the user's identity construction on social media.

Aspects of multilingualism, learner Englishes and English as a lingua franca guide Lorenz's investigation of the discourse marker *like* in the United Arab Emirates. As part of a larger project on *Language repertoires and attitudes of students in the United Arab Emirates* (see, e.g., Siemund et al., 2021), her study tests various potential factors that might influence the occurrence of *like*.

Set in the related context of Saudi Arabia, AlShurfa's et al. study is an example of the contact type of Learner Englishes. The authors discuss a range of syntactic features, demonstrating how transfer from L1 Arabic influences the shape of Saudi English.

Learner Englishes and the importance of English as a global, international language is at the center of Peterson's article on English in Finland. The author takes a critical view as, on the one hand, English has become part of Nordic Exceptionalism, which is expressed in the widely held and misguided belief that "everybody can speak English." On the other hand, the teaching of English is still strongly geared towards so called native-speaker norms, perpetuating widely criticized imbalances across the Englishes-speaking world (cf. Kirkpatrick, 2021 for a general discussion of this issue).

Finally, two articles in the Research Topic address aspects of Irish English that fits the classification of Koiné Englishes from a current point of view. Shimada traces the history of how Irish or Hiberno English was formed as a language shift variety in a bilingual constellation before focusing on three grammatical constructions that exemplify Irish contact and that have been retained as characteristic features of Hiberno English on its way of having turned into an L1 variety shaped by dialect contact over the past century.

Kirk takes the discussion of Irish English onto a more general level of world Englishes. Drawing from different data sources (ICE corpora, handbooks, the electronic World Atlas of Varieties of English, and the GloWbE corpus), he investigates how Irish English fits into models of world Englishes.

The different facets of language contact involving Englishes addressed in this Research Topic emphasize the diverse manifestations and influences of contact between Englishes and other languages. While this diversity can be organized in line with the Language Contact Typology summarized and applied here, all the contributions investigate unique aspects within these broad types. Moreover, the articles draw from a wide range of data, such as social media language, corpora, elicited responses, questionnaires, and interviews, and they provide further insights into the fundamental role of language contact for the uses and functions of Englishes worldwide.

Author contributions

AO and PS jointly planned and drafted the article. AO completed and finalized the piece. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Blommaert, J. (2010). *The Sociolinguistics of Globalization*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511845307

Kirkpatrick, A. (2021). "Teaching (about) world englishes and english as a Lingua Franca," in *Research Developments in World Englishes*, ed A. Onysko (London: Bloomsbury Academic), 251–270. doi: 10.5040/9781350167087.ch-012

Onysko, A. (2016). Modelling world Englishes from the perspective of language contact. *World Englishes* 35, 196–220. doi: 10.1111/weng.12191

Siemund, P. (2022). *Multilingual Development: English in a Global Context*. Cambridge: Cambridge University Press.

Siemund, P., Al-Issa, A., and Leimgruber, J. (2021). Multilingualism and the role of English in the United Arab Emirates. *World Englishes* 40, 91–204. doi: 10.1111/weng.12507

Siemund, P., and Leimgruber, J. (2021). *Multilingual Global Cities: Singapore, Hong Kong, Dubai*. Singapore: Routledge.

The Multilingual Pragmatics of New Englishes: An Analysis of Question Tags in Nigerian English

Michael Westphal *

English Seminar, University of Münster, Münster, Germany

This paper presents a variational pragmatic analysis of multilingual question tags in Nigerian English, combining a corpus-pragmatic analysis of the Nigerian component of the International Corpus of English with a survey study on the preferences and attitudes of Nigerian students toward different question tag forms. The corpus study highlights multilingual pragmatic variation in terms of form and function of variant as well as English and non-English (i.e., derived from indigenous Nigerian languages) invariant question tags in six text types: conversations, phonecalls, classroom lessons, broadcast discussions, broadcast interviews, and legal cross-examinations. Nigerian speakers combine a wide range of English and non-English invariant forms, whereas variant question tags only play a marginal role and are not characteristic of Nigerian English. Text type influences the overall frequency of question tags and – together with the pragmatic function – constrains the use of individual forms. The survey study shows diverging results as the participants generally prefer variant over invariant question tags and show a strong dispreference for indigenous Nigerian forms when speaking English. Nevertheless, their preferences for specific forms over others are guided by the communicative setting and requirements of a given situation. The students also hold most positive attitudes toward variant question tags, while non-English tags are rated less positively on items reflecting decency. However, all question tag forms are valued in terms expressiveness. Hence, Nigerian students' dispositions toward multilingual question tag use are guided by a prescriptive ideology that is biased toward canonized English forms. While indigenous Nigerian forms are well integrated into question tag use, indicating a high degree of nativization of Nigerian English at a pragmatic level, acceptance for these local forms is lagging behind. In general methodological terms, the paper shows that question tags – or discourse-pragmatic-features in general – have high potential for studying multilingual variation in New Englishes. However, studies on the multilingual pragmatics of New Englishes need to consider the full range of multilingual forms, take into account variety-internal variation via text type, and should ideally also study the users' perspectives.

Keywords: Nigeria, New Englishes, multilingualism, International Corpus of English, corpus pragmatics, variational pragmatics, attitudes

INTRODUCTION

The World Englishes paradigm has pushed the decolonization of the academic study of the English language by highlighting the global diversity of the English language along different national Englishes (e.g., Kachru, 1985). Much research in this area has focused on New Englishes, which are varieties of English that have developed out of colonial contact situations. New Englishes are used in countries where English serves an official function but is usually learned as a second/subsequent language (or dialect). The notion of New Englishes was first used to describe emerging national standard varieties (Platt et al., 1984), such as Indian or Nigerian English, but may also include English-based creoles (Mufwene, 1994) or hybrid Englishes (Schneider, 2016). All of these different types of New Englishes exist in highly multilingual ecologies and monolingualism is most often the marked case. For example (Anderson and Ansah, 2015: 60), describe code-switching as so pervasive in many domains of language use in West Africa that it has become the norm rather than an act of identity. Consequently, the development of Standard New Englishes, which emerge in such multilingual environments and are not fully codified, can be assumed to be strongly affected by other languages and dialects.

A case in point is Nigerian English, the emerging Standard English of Nigeria (Schneider E., 2007: 199–212). Nigeria is located in West Africa and is home to more than 140 million inhabitants from diverse ethnic groups who also speak various languages/dialects (Jowitt, 2019: 4–5). As the distinction between language and dialect is very complex, estimations range from around 400 to more than 500 languages in Nigeria. The three major ethnic groups and their first language are Hausa, Yoruba, and Igbo, which are recognized officially alongside English. English is mainly learned as a second/subsequent language and functions as an interethnic lingua franca in many formal domains, such as the government or higher education. Nigerian English is not a monolithic fixed norm but exhibits a high degree of variation, for example in terms of ethnicity or the level of education of speakers (Jowitt, 2019: 24–33). In addition to the different first languages, Nigerian English is also influenced by Nigerian Pidgin, an English-based contact language that is mainly used as a lingua franca in informal domains, such as on the market or in public transportation (Deuber, 2005). The distinction between these two varieties is increasingly blurred as there is mutual borrowing and Pidgin has been making inroads into domains formerly reserved for Nigerian English, such as the media (Schneider E., 2007: 207–209). In support of the codification of Nigerian English (Gut, 2012), many studies on this variety have focused on describing its grammatical (e.g., Gut and Fuchs, 2013), phonetic (e.g., Oyebola et al., 2019), and pragmatic (e.g., Unuabonah and Gut, 2018) properties. Much of the most recent research on Nigerian English has used the Nigerian component of the International Corpus of English (ICE) (Wunder et al., 2010).

The ICE project (Greenbaum and Nelson, 1996) includes national corpora of Standard Englishes from countries where English has an official status. Each corpus has a size of one million words and has the same design, covering 15 spoken and 13 written text types. Hence the ICE corpora are a rich tool for analyzing variation across and within World Englishes and has been used extensively (e.g., Hundt and Gut, 2012). The ICE corpora are designed to represent Standard English exclusively, which means that multilingual/-dialectal variation is often suppressed in the compilation process. This bias toward monolithic standard language ensures comparability across corpora but fails to depict the actual multilingual embedding of Standard New Englishes. Hence (Mair, 2011: 234), states that the failure to recognize the embedding of New Englishes in intensely multilingual communities is the most "glaring lacuna" in corpus-based research on New Englishes and he argues for the compilation of multilingual corpora. However, recent corpus-pragmatic research has demonstrated the presence of many 'non-English' discourse-pragmatic-features (i.e., syntactically optional particles used to express stance, to guide utterance interpretation, or to structure the discourse; Pichler, 2013: 4) in ICE-Nigeria, such as *abi*, *o*, or *sha* (e.g., Unuabonah and Oladipupo, 2018). Hence, discourse-pragmatic-features seem to be a promising area to analyze multilingual variation in New Englishes.

In this paper, I highlight multilingual variation in Nigerian English by analyzing the use and perception of question tags, which I treat as a set of discourse-pragmatic-features that includes English (e.g., *isn't it*, *right*) and non-English/indigenous Nigerian forms (e.g., *abi*, *o*). On the one hand, I analyze the use of question tags in ICE-Nigeria, demonstrating that the ICE corpora can be used to study multilingual variation in New Englishes. This corpus-pragmatic analysis highlights internal variability in Nigerian English by investigating the use of multilingual question tags across six dialogic text types: conversations, phonecalls, classroom lessons, broadcast discussions, broadcast interviews, and legal cross-examinations. On the other hand, I investigate Nigerian students' perception of multilingual variation of question tag use in a survey study. The survey includes a multiple-choice task, in which the participants indicate which question tag form they prefer in different situations, and a written Matched-Guise-Test, in which the participants rate the use of different question tag forms on attitudinal scales. With this mixed-methods approach, I address the following research questions:

- Which question tag forms do Nigerian speakers use when speaking English?
- How does text type influence the overall distribution of question tag forms?
- How do text type and function constrain the selection of particular forms over others?
- Which question tag forms do Nigerians prefer in different situations?
- Which attitudes do Nigerians hold toward different question tag forms?

This paper makes an important contribution to the description of Nigerian English and highlights new methodological paths to using the ICE. I also present new methods to analyze speakers' perception of discourse-pragmatic-features. On a theoretical level, the paper

merges a fundamental assumption of sociolinguistics with multilingualism in World Englishes. Thus, I argue that the structure of New Englishes can only be understood by illustrating their internal variation, which in large parts is caused by their multilingual embedding. Question tags are used as a case in point to illustrate the structure of multilingual variation in Nigerian English.

The paper is structured as follows. **Section 2** discusses previous variational-pragmatic research on World Englishes, focusing on previous studies on question tags. In **Section 3**, I present the methods of the corpus-pragmatic and the survey study. The findings of these two studies are presented in **Sections 4 and 5**. **Section 6** discusses these findings and highlights the methodological implications for research on multilingual variation in New Englishes.

PRAGMATIC VARIATION IN WORLD ENGLISHES

Most descriptive (Kortmann and Schneider, 2008) and comparative work (e.g., Hundt and Gut, 2012; Siemund, 2013) in World Englishes has focused on morpho-syntax, lexicon, and phonetics. ICE-based research has largely focused on the first two levels of variation, but newer ICE corpora also allow studying phonetics (e.g., Oyebola et al., 2019). The results of ICE-based research on New Englishes are mostly compared to Englishes spoken as a native language, mainly British English, to delineate their level of nativization based on the degree of difference. See (Hansen, 2018: 48–54) for a critical discussion of this approach.

Pragmatic phenomena, including discourse-pragmatic-features, have been studied less frequently. Traditionally, they are not used as indicators of nativization. Instead their use is ascribed to idiosyncratic preferences. For example (Bautista, 2011: 81), describes the use of 'no by Filipino teachers as a verbal tick of some speakers. Aijmer (2013: 145) explains the use of discourse-pragmatic-features by speakers of New Englishes as a result of their lower competence in English. In his overview of African Englishes (Mesthrie, 2008: 30), dismisses invariant question tags as "garden-variety" structures of New Englishes.

Pragmatics, as the study of language in context, mirrors the research gap in World Englishes as regional and social variation has been neglected. Schneider (2012: 464) illustrates that pragmatics was initially concerned with establishing seemingly universal theories of speech acts, politeness, or the structure of conversation, but conventions in and across particular languages/varieties did not play a role. The discipline of variational pragmatics (Schneider and Barron, 2008) fills this research gap by investigating pragmatic variation with regard to region and other macro- (e.g., social class) or micro-sociolinguistic (e.g., power) factors. Variational pragmatics mostly draws on corpus or survey data. Corpus-pragmatic analyses often combine qualitative coding of individual pragmatic phenomena (i.e., horizontal reading of corpus texts) with quantification (i.e., vertical reading of corpus texts) (Aijmer and Rühlemann, 2015: 3–9). In addition, research on speech acts in different varieties has shown the benefits of combining analyses of

language use with survey data (e.g., Schneider K. P., 2007). Besides speech acts, the pragmatic phenomena most often studied in this field of research are discourse-pragmatic-features. However, there has been a strong focus on English varieties spoken as a native language, such as British English (e.g., Pichler, 2013; Beeching, 2016). In addition, corpus analyses of discourse-pragmatic-features have mainly focused on face-to-face conversation, while other text types have been neglected. Moreover, there are hardly any studies on discourse-pragmatic-features that have utilized survey data; an exception is Beeching (2016). Survey data in general is rare for studying pragmatic phenomena in New Englishes; exceptions include Schröder and Schneider (2018) and Anchimbe (2018).

If discourse-pragmatic-features are studied in New Englishes, for example in Nigerian English, the full range of multilingual variation for these pragmatic phenomena is not considered. On the one hand, research focuses on individual 'non-English' forms, which derive from Hausa, Igbo, or Yoruba and are integrated into Nigerian English – often via Nigerian Pidgin. Unuabonah and Oladipupo (2018) investigate *abi*, *o*, and *sha*, Unuabonah and Oladipupo (2021) analyze *jare*, *biko*, *jor*, *shebi*, *shey*, and *fa*, and Unuabonah (2020) examines *na wa*, *shikena*, *ehn*, and *ehen*. There are also studies on indigenized uses of English forms. Oladipupo and Unuabonah (2021) analyze the particle *now* in Nigerian English. All these studies use ICE-Nigeria, show the general frequencies of these forms, and list their different pragmatic functions. However, they do not investigate the sociolinguistic dynamics of pragmatic variation, for example by analyzing the constraints of use of the different discourse-pragmatic-features (e.g., text type, age, or gender). In addition, the local discourse-pragmatic-features are not analyzed in relation to alternative English forms which may fulfill similar pragmatic functions.

If entire sets of discourse-pragmatic-features are studied, then the focus is on English forms and their use is compared to British English. Unuabonah and Gut (2018) investigate 173 commentary pragmatic markers, Unuabonah (2019) analyzes 71 discourse markers, and Unuabonah et al. (2021) examine the 64 intensifiers in ICE-Nigeria. The concordance lists of discourse-pragmatic-features are based on previous research on English varieties spoken as a native language and do not include non-English forms. These studies generally conclude that Nigerian English shows an overall lower frequency of the selected discourse-pragmatic-features but there are distinct patterns of use. According to these studies, speakers of Nigerian English use a reduced inventory of discourse-pragmatic-features and demonstrate a lower stylistic variability in comparison to British English. These conclusions seem somewhat biased as indigenous forms are not included. Despite this already large and still growing body of research on the pragmatics of Nigerian English, there is the need to investigate multilingual variation for an entire set of discourse-pragmatic-features that includes English and non-English forms and considers constraints of variation (e.g., text type). This approach allows expanding the understanding of the structure of Nigerian English and by extension of New Englishes in terms of the dynamics of multilingual variation.

Question tags are one set of discourse-pragmatic-features frequently studied in World Englishes. There is a wide range of forms that can function as question tags and hence they are not defined by their form but by their function. In most general terms, speakers append question tags to statements to receive a confirmation from their interlocutors, to integrate other participants in the conversations, or to emphasize their statements (Wilson et al., 2017: 732–734; Kimps, 2018: 14–27). In terms of forms there is a major distinction between variant (also called canonical) and invariant question tags. Grammars (e.g., Biber et al., 1999: 208–210) mostly focus on the formal properties of variant question tags, whose structure depends on the main clause they are attached to.

Variant question tags consist of a pronoun and an auxiliary verb, which is identical to the one in the main clause 1); if there is no auxiliary verb in the main clause, *do* is used 2). The main clause and the question tag also agree in terms of tense, aspect, and mood (1–2). However, the use of these forms often does not align with these rules. For example, many speakers commonly use invariant *isn't it* 3). Canonical question tag constructions are a typological anomaly, which is typical for English and only found in few other languages (see Axelsson, 2011: 823–829).

1) <#>But I should take it again <#>But I wasn't dictating *was I* (les_13)
2) <#>You like it *don't you* (con_09)
3) <#>We'll be selling them for six thousand *isn't it* (ph_01)

Invariant question tags have a fixed form that does not depend on the main clause to which they are attached, and they are not discussed in much detail in grammars (e.g., Biber et al., 1999: 1,089). Invariant question tags may be single words or particles, such as *right* or *eh*, as well as multi-word units, such as *you know* (4–6).

4) <#>It's not like as if he's the Messiah *right* (con_09)
5) <#>Is no is no good kuli it was not the good one *eh* (con_45)
6) <#>He just cannot condone such hypocrisy *you know* (ph_01)

In addition to such English forms, invariant question tags also include forms borrowed from indigenous Nigerian languages, such as *sha* or *abi* (7–8), which are mainly used to add emphasis in these two examples.

7) <#>No I don't really like chilled water *sha* (con_36)
8) <#>Philosophy is not maths at all *abi* (con_50)

Previous research on question tags in World Englishes exhibits several gaps. The overwhelming amount of studies on question tags has been done for Englishes spoken as a native language (e.g., Tottie and Hoffmann, 2006; Gómez González, 2018) and there is a dearth of research on other text types than conversations (see Barron, 2015: 224). Furthermore, most research has focused on variant question tags exclusively. For example, Borlongan (2008), Parviainen (2016), and Hoffmann et al. (2017) analyze variant question tags in Asian Englishes comparing them to British (and American) English. All three studies find a very high number of

invariant uses of *isn't it*, *is it*, or *is it not* and they conclude that these invariant uses are characteristic for Asian Englishes. However, this conclusion seems biased as invariant question tags were not considered. A notable exception is Columbus (2009, 2010), who studies invariant question tags across several Englishes and shows that specific question tag forms are often typical for individual varieties, such as *na* for Indian English or *eh* for New Zealand English. She also illustrates that a wide range of forms which are often not included in analyses and descriptions of question tags are highly frequent across varieties, such as *OK*, *yeah*, *you know*, or *you see*. Similarly, Takahashi (2014) shows that speakers of Asian Englishes combine various English and indigenous question tag forms. Research on discourse-pragmatic-features in Singaporean English and Singlish (see Leimgruber, 2013: 84–96), such as *ah*, *meh*, or *lah*, does not discuss these forms as question tags nor in relation to English alternatives.

Gómez González (2018) investigates both types of question tags in British English, showing that variant ones are five times as frequent as invariant ones. In contrast, recent research on New Englishes (Wilson et al., 2017; Mbakop, 2020; Westphal, 2020) has shown that invariant question tags outnumber variant ones by far. These studies have also demonstrated that the use of *isn't it* is rare. In their analyses of Trinidadian and Philippine English, respectively, Wilson et al. (2017) and Westphal (2020) also show that text type exerts a strong influence on the general frequency of question tags and on individual forms. For example, non-English forms in New Englishes are more common in informal text types, or *OK* is most commonly used in classroom lessons by teachers. Both studies demonstrate that individual question tags serve specific functions. For example, *eh* is mainly used to add emphasis and variant question tags are preferentially used to receive confirmation from interlocutors.

Despite this prevalence of invariant forms in New English, there seems to be strong prescriptivism in terms of question tag usage in New Englishes contexts, such as anglophone West Africa. Mbakop (2020: 1) argues that in Cameroon English Language Teaching focuses very strongly on variant question tags, which is in strong contrast with usage patterns of Cameroonian English. One of the sketches of the Ghanaian-American comedian Ebaby Kobby makes fun of the 'incorrect' use of question tags among Ghanaian students.[1] There are also prescriptive papers on the correct use of question tags in Nigerian English (Osakwe, 2009). Hence, there seems to be a strong standard language ideology which devalues invariant question tags, but so far, no perception study on English question tags has been carried out.

DATA AND METHODS

The analysis of the use of question tags in Nigerian English utilizes the spoken component of ICE-Nigeria (Wunder et al., 2010), which was published in 2015 and includes transcriptions

---

[1]See online at: https://www.youtube.com/watch?v=trLjM5XHELo.

TABLE 1 | Dialogue subcorpus from ICE-Nigeria.

Text type	Text code	Number of texts	Word count
conversations[a]	con	30	87,225
phonecalls	ph	7	21,310
classroom lessons	les	14	42,187
broadcast discussions	bdis	26	41,313
broadcast interviews	bint	10	20,744
legal cross-examinations	cr	10	20,973

[a]Only a random selection of 30 conversations was used for the analysis.

and sound files for most texts. Like other ICE-corpora the spoken component has a size of 600,000 words but individual texts vary in size, in contrast to the standardized ICE format of 2,000 words per text. Individual text types have the same subcorpus size as in other ICE-corpora. As question tags are an integral part of spoken dialogues and more common in these text types than in others (e.g., Borlongan, 2008), this corpus-pragmatic study of ICE-Nigeria only uses texts from six dialogue text types: conversations, phonecalls, classroom lessons, broadcast discussions, broadcast interviews, and legal cross-examinations.[2] The dialogic subcorpus analyzed in this paper includes 97 texts with an overall size of 233,752 words. **Table 1** shows the number of texts and the word count for each text type.

These six text types differ substantially in terms of the communicative setting, which includes the level of formality, the degree of prestructuring of the dialogue, and speakers' roles in the given discourse. Conversations and phonecalls are private dialogues, they are the least formal, open in terms of their structure, and the interactions are very diverse as speakers do not have fixed roles. This is different to the four public dialogues, which are all more formal, have a higher degree of prestructuring, and speakers fulfill specific roles. In classroom lessons, there is one teacher/lecturer who gives explanations and asks questions to their students. In broadcast discussions and interviews, hosts moderate these public dialogues and guests answer the hosts' questions, voice their opinion, or debate with each other. Legal cross-examinations are the most formal text types. They are rigidly structured and controlled, as attorneys question witnesses, who have to testify in front of court.

Due to the form-function mismatch of question tags, a top-down concordance analysis was not possible. Searching for a specific form that may function as a question tag produces numerous concordances that are not question tags. For example, *right* may be used as an adjective (e.g. *the right choice*), to backchannel, or as a question tag. In addition, relying on a pre-defined list of question tag forms may lead to biased results. There might be many forms that fulfill the function of question tags but are not included in this list and are thus overlooked.

Hence, I read and – if the sound files were available[3] – listened to all 97 texts. I identified and coded each question tag token

qualitatively. Much previous corpus-based research on question tags has mainly relied on formal criteria to define a question tag. Many studies have analyzed variant question tags exclusively (e.g., Tottie and Hoffmann, 2006; Borlongan, 2008; Barron, 2015; Parviainen, 2016; Hoffmann et al., 2017), while others also allow invariant forms but only focus on sentence-final question tags (e.g., Takahashi, 2014). Instead, I applied a much wider understanding of question tags similar to Columbus (2009, 2010) and tried to capture all forms that may function as a question tag, using function-based criteria to define what counts as a question tag. Pichler (2010, 2013: 28–32) discusses the problems of such an approach for variationist analyses of discourse-pragmatic-features. In the conclusion, I revisit this methodological issue.

For the purpose of the current study, the decision whether a specific form functions as a question tag is grounded on the following criteria. Question tags are discourse-pragmatic-features (i.e., they are syntactically optional), and they are attached to utterances. Question tags are neither fillers (i.e., forms surrounded by repetitions or other fillers, such as *uhm*, were excluded) nor entire utterances on their own (e.g., *right* used as a backchannel). Furthermore, they are neither items used in their full literal sense (e.g., *the right choice*) nor part of fixed expressions (e.g., *right now*). In cases of repetitions of a question tag in an utterance, only the final form was counted. The main functional criterium is that question tags fulfill an informative, facilitative, or punctuational function. A form was identified as a question tag if it fulfills one of these functions.

Speakers use question tags in an informative way when they are unsure of the content of an utterance, and they want new information or a confirmation for the assumption they have expressed. An answer is expected when question tags are used informatively. In (9), the attorney demands information whether the witness has written a piece of information themselves. Speakers use question tags in a facilitative way to integrate interlocutors more into the discourse either by signaling that they are willing to hand over their turn or to invite (verbal or non-verbal) responses. In (10), a teacher adds *OK* to their utterance to check whether the students have understood the explanation and invites them to backchannel or to interrupt if there are any uncertainties. Question tags with a punctuational function are used for stylistic purposes, mainly to add emphasis. The speaker is sure about the content of the utterance and mostly no answer is expected. In (11), the speaker uses *o* to add emphasis to his suggestion to his interlocutors.

9) <#>Did you write it yourself *not so* (cr_10)
10) <#>This times four is four *OK* (les_06)
11) <#>You guys should call him before he goes *o* (con_11)

This tripartite functional distinction is a reduced system of previous functional classifications (e.g., Algeo, 1990; Tottie and Hoffmann, 2006; Wilson et al., 2017). A six-way distinction into confirmatory, facilitating, attitudinal/punctuational, informational, peremptory, and aggressive tags, which is based on Algeo (1990), has been widely used but proved problematic for the Nigerian data. A distinction between informative and confirmatory uses, which is

---

[2]Parliamentary debates and business transactions were not used as discourse-pragmatic-features are rare in the former text type and texts from the latter type are very heterogenous in terms of the communicative setting.

[3]Sound files were available for all texts except for four conversations: con_9, con_11, con_12, and con_16.

based on different levels of prior knowledge of the speaker (Tottie and Hoffmann, 2006: 299), proved impossible to make. Hence, these two functions were merged to one category, which I labelled informative. Similarly, peremptory and aggressive tags could hardly be distinguished from general attitudinal ones. Thus, these three categories were fused to one, which I labelled punctuational. Tottie and Hoffmann (2006) and Wilson et al. (2017) classify over 90% of their question tags as confirmatory, facilitating, and attitudinal/punctuational. Thus, the current three-way distinction (i.e. facilitative, informative, punctuational) covers the main uses, facilitates coding and allows using function more easily in regression modelling. However, there are also cases in which it was very difficult to clearly distinguish between different pragmatic functions, and as question tags are multifunctional, they sometimes fit into more than one category. In these cases, question tags were ascribed to all the respective functional categories and were coded as multifunctional. In (12), the guest (<$A>) in a broadcast discussion uses *you see* both to add emphasis to his argument but also invites backchannelling from the host (<$B>) and the other guest, checking whether they are still following his line of argumentation.

12) <$A><#>Let me tell you why we cannot do anything for now *you see* <$B><#>Mhm (bdis_12)

These codings of the forms and pragmatic functions were used for a quantitative analysis of the use of question tags in the six dialogue text types from ICE-Nigeria. For all statistics on text type variation, conversations and phonecalls were merged to the overarching category of private dialogues. Broadcast discussions and interviews were unified to the single category broadcast dialogues. Normalized frequencies are presented as tokens per fifty thousand words (tpf). The descriptive statistics first describe the general diversity in question tag forms in the subcorpus. Second, I demonstrate the overall frequency distribution of variant vs. invariant question tags across the different text types. Third, I highlight the distribution of the most frequent English and non-English question tag forms across the text types and their functional diversity. I define non-English forms as forms borrowed from indigenous Nigerian languages (as defined in previous research: e.g., Unuabonah and Oladipupo, 2018; Jowitt, 2019; Unuabonah, 2020; Unuabonah et al., 2021).[4] English forms may also include local innovative forms unusual for many varieties of English, such as *not so* or *no be*.

The inferential statistics then show in detail how text type and function constrain the use of selected question tag forms: *o*, *OK*, *you know*, and variant question tags. These forms were selected as they are sufficiently frequent across the corpus and serve to illustrate different socio-pragmatic profiles of question tags in Nigerian English. For this analysis, I used binary regression models in RBRUL (Johnson, 2009) with form as a dependent variable, which was reduced to a binary distinction. In one model, the use of one form is compared to all others (e.g., *o* vs. all other

forms). Text type (four levels: private dialogues, classroom lessons, broadcast dialogues, legal cross-examinations) and function (three levels: facilitative, informative, punctuational) were used as fixed predictor variables. Speaker was inserted into each model as a random factor to avoid Type I errors as idiosyncratic variation is very pronounced for individual question tag forms (e.g., Wilson et al., 2017: 734). Multifunctional question tags were excluded from the regression analyses. For all binary regressions *p*-values for the general effect of a factor and centered factor weights for the individual levels are given. Factor weights indicate the direction and size of the effect. They range from 0 to 1; values above 0.5 signal a preference and values below 0.5 a dispreference.

The survey study combines a multiple-choice task, which analyzes the participants' preferences for different question tag forms in various scenarios, and a written Matched-Guise-Test, which investigates participants' attitudes to a range of question tag forms.[5] The multiple-choice task is modelled after a discourse-completion-task but with fixed answer options. This means that there is a brief description of a dialogue and the participants had to imagine that they are a part of this situation. This communicative scenario is followed by a dialogue, which the participants completed by selecting a question tag form from pre-given options. There are seven scenarios, which are based on corpus data, reflect different text types, speaker roles, and speaker relationships (e.g. teacher-student, attorney-witness). In addition, the scenarios target a specific pragmatic function of question tags (facilitative, informative, punctuational). For each scenario the participants chose from ten different options including non-English forms (*abi*, *o*, *sha*), English forms (*eh*, *OK*, variant question tag, invariant *isn't*, *you know*), and the option 'other', which participants could fill out freely.

- Scenario 1 is an informal private dialogue between friends and targets an informative question tag.
- In scenario 2, participants imagine that they are a teacher in a classroom lesson and use a question tag facilitatively to check whether students have any questions.
- Scenario 3 depicts a legal cross-examination. Participants imagine they are an attorney eliciting information from a witness using an informative question tag.
- Scenario 4 is an informal private dialogue between friends and requires the use of an informative question tag.
- Scenario 5 is a classroom lesson. Participants assume the role of the teacher adding a punctuational question tag to an imperative targeted at a student.
- Scenario 6 is an informal private dialogue between friends, which targets the use of a facilitative question tag.
- In scenario 7, participants imagine they are a student at university and ask their lecturer for a confirmation of information using a question tag.

The second part of the survey investigates the participants' attitudes toward the same nine forms given as options in the

---

[4]See Jowitt (2019: 139–141) for a discussion of the origin and meaning potential of several indigenous discourse markers typical of Nigerian English.

[5]The survey can be accessed online at: https://www.uni-muenster.de/imperia/md/content/englischesseminar/question_tag_survey_-_nigeria.pdf.

multiple-choice task, using a modified Matched-Guise-Test with written stimuli, adapted from Beeching (2016: 38–41). The participants were presented with nine scenarios that contain a description of the situation which is typical of the pragmatic profile of the question tag form under investigation as used in ICE-Nigeria. The description is followed by two almost identical utterances produced by two speakers. One contains a question tag and the other does not. The participants were asked to rate the speaker who uses the question tag in contrast to the other speaker on six personality traits, using six-point semantically differential scales (impolite vs. polite; reserved vs. outgoing; aggressive vs. gentle; indirect vs. direct; unfriendly vs. friendly; uneducated vs. educated). Hence, participants' attitudes toward question tag forms were elicited indirectly (see Garrett, 2010: 39–43). An additional blank space was given, where participants could add their additional thoughts on the use of the specific question tag form.

The descriptive statistics for the multiple-choice task describe the frequencies of selection for each scenario individually. For individual question tag forms, binary regression models were run to highlight how scenario influences the choice of one form over the others. Form was used as the dependent variable (e.g., variant question tag vs. all other options) and scenario (7 levels) was the fixed predictor variable. The descriptive statistics for the Matched-Guise-Test illustrates the ratings of all nine forms on the six scales. Principal Component Analysis was used to investigate how the different items pattern together, illustrating the underlying attitudinal dimensions (Garrett, 2010: 55–56). To illustrate the ratings of the nine question tag forms along the different dimensions, mean values of all items that clustered together were calculated for each participant. These mean values were then used in further regression analyses. Rating scores for the different dimensions were employed as dependent variables and question tag form was used as fixed predictor variable.

University students were selected as the target demographic for the survey. Students might be a convenience sample but as they are generally fluent in English, young, well-educated, and hence part of the future middle-class, they are an important group for the emerging standard variety in Nigeria. Fieldwork was carried out by Folajimi Oyebola at the University of Lagos in January 2020 among 1st year students who participated in the class *Use of English in Nigeria*. This is a compulsory class students across all disciplines must attend in their first year at the University of Lagos. 49 Nigerian students completed the questionnaire. The larger majority of them is female (32 female; 16 male; 1 no answer) and between 18 and 25 years old (41 18–25; 5 younger than 18; 3 no answer). Of the students who indicated their type of studies ($N = 40$), the majority studied law (18; 45.0%), followed by 'English' (13; 32.5%),[6] adult education (7; 17.5%), and educational administration (2; 5.0%). The sample is rather small and does not allow any internal

differentiation but gives first insights into the perception of different question tag forms in Nigerian English.

RESULTS I: CORPUS-PRAGMATIC ANALYSIS OF DIALOGUES IN ICE-NIGERIA

Question tags are a frequent feature in the six dialogue text types of ICE-Nigeria. 1,326 tokens (284.85tpf) were identified in the 97 texts. Nigerian speakers use a wide range of English and non-English forms. **Table 2** provides an overview of these different forms. The overall most frequent question tag form is *you know* followed by *now*, *OK*, and *o*. With only 33 occurrences, variant question tags are marginal. Of these 33 question tags only 15 agree with the main clause they attach to in terms of the auxiliary (or use *do* in a canonical way), tense, aspect, and mood. Of the 18 invariant uses, 13 occurrences are invariant *isn't it* or *is it* (*not*) as in (3) and (13), but there are also other invariant uses as in (14).

13) <#>No but you're talking about the G twenty summit *isn't it* (bdis_01)
14) <#>You attended University of Illorin *don't you* (con_06)

The speakers of Nigerian English from the subcorpus use a mix of English and non-English forms. 832 tokens (62.7%) were classified as English, while 261 (19.7%) question tags derived from indigenous Nigerian languages or Nigerian Pidgin. The most frequent indigenous forms are *o*, *abi*, and *sha*. *Now* accounts for 233 question tag occurrences. However, *now* is a special case as there is variation in spelling between *na*, *ne*, and *now*. The most frequent spelling is *now* (216), followed by *na* (13), and *ne* is least frequent (4). These spellings do not match the variation in pronunciation consistently, which includes [naʊ], [naᵘ], [na(:)], [nə], [næ], and [nɛ], the latter two being very rare. In addition, there is variation in the tonality of *now* (Oladipupo and Unuabonah, 2021: 375–377). The extended pragmatic uses of *now* in Nigerian English can be viewed as a result of nativization of the general English pragmatic particle *now* (Oladipupo and Unuabonah, 2021), however, the Nigerian Pidgin form *na* might overlap with *now* (Unuabonah et al., 2021). Whereas previous studies have looked at *now* and *na* independently (Oladipupo and Unuabonah, 2021; Unuabonah et al., 2021), a detailed analysis that looks at *now*, *na*, and *ne* is required. In addition, this analysis should also study the exact phonetic realization *now* and its spelling variants. Due to this complexity of *now* a separate analysis is needed and *now* is not discussed further in this paper.

Text type has a strong influence on the frequencies of question tags in the dialogue subcorpus. **Figure 1** shows the normalized frequencies of variant and invariant question tags for the four general dialogue text types. Question tags are most frequent in private dialogues closely followed by classroom lessons. In broadcast dialogues, question tags are less than half as frequent as in private dialogues. Question tags are by far the least frequent in legal cross-examinations. Hence, a general correlation with formality becomes apparent. The more formal the text type the fewer question tags are used. Despite this general effect of text type on question tag frequencies, there is also

---

[6]The category 'English' includes a range of answers referring to English in one way or another, such as "education (English)", "English language", or "English". However, it seems likely that some students misunderstood the question and inserted the name of the class.

TABLE 2 | Overview forms.

English forms		Non-English forms	
Form	**Frequency (%)**	**Form**	**Frequency (%)**
alright	35 (2.6)	*abi*	31 (2.3)
eh	68 (5.1)	*ah* (i.e. *a*, *ah*, *a-a*)	17 (1.3)
OK	151 (11.4)	*ba*	11 (0.8)
right	37 (2.8)	*o*	140 (10.6)
huh	16 (1.2)	*sef*	12 (0.9)
variant question tags	33 (2.5)	*sha*	44 (3.3)
yeah	18 (1.4)	other non-English (*ha*, *nko*, *na so*, *oya*)	6 (0.5)
yes	18 (1.4)		
you see	33 (2.5)	**mixed forms**	
you understand	46 (3.5)	**form**	**frequency %**
you know	330 (24.9)	*now* (i.e. *now*, *na*, *ne*)	233 (17.6)
other-English (e.g., *correct*, *no*, *no be*, *or*, *or not*, *or what*)	47 (3.5)		

FIGURE 1 | Overall question tag frequencies across the four text types in ICE-Nigeria.

FIGURE 2 | Text type variation of individual question tag forms (based on normalized frequencies).
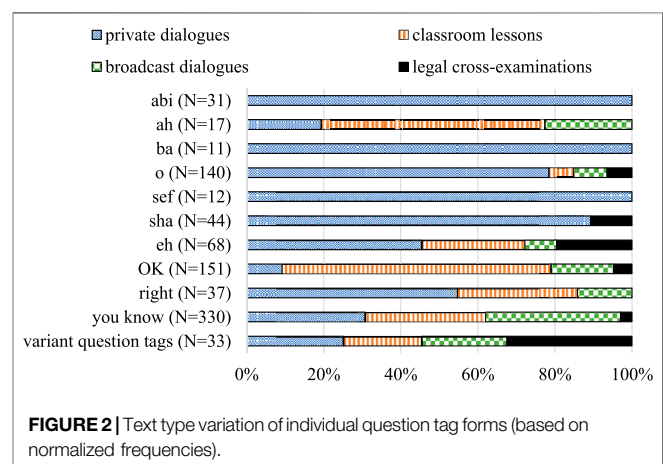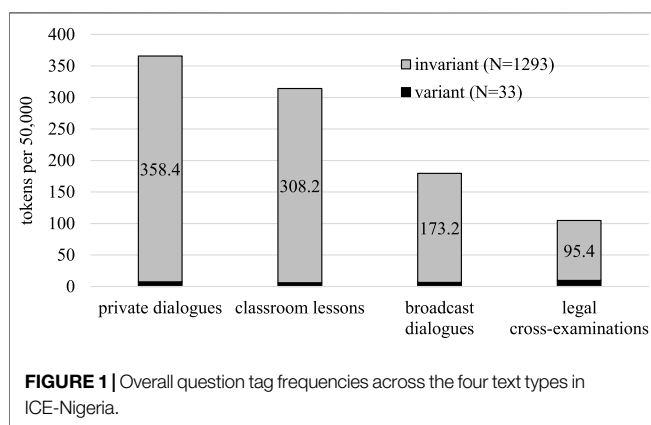
substantial internal variation for text type. This is most pronounced for classroom lessons. For example, in les_11 the teacher does not use any question tag and for les_01 (105.6tpf), les_05 (113.5tpf), and les_07 (127.6tpf) normalized frequencies are very low. In contrast, the teachers in les_06 (454.8), les_03 (770.0tpf), and les_14 (832.1tpf) use a very high frequency of question tags. This variation has to do with the different teaching styles. Teachers with low frequencies of question tags often dictate entire passages to their students, hence these texts have a quasi-written character. Les_05 and les_07 are special cases as these texts are bible classes and all participants read out passages from the bible extensively with very little dialogue between them. In contrast, teachers who use many question tags rely on oral explanations, want to make sure their students are following along, and interact more directly with them. Hence, idiosyncratic variation in terms of question tag use is guided by the exact type of classroom lesson and the pedagogical style of the teacher.

Text type also has a substantial effect on the distribution of individual English and non-English question tags, which is shown in **Figure 2**. Generally, non-English question tags are mostly used in private dialogues, and they are rare or even completely absent in public and more formal text types. This text type variation is categorical for *abi*, *ba*, and *sef* and very pronounced for *o* and *sha*. *Ah* is an exception as this form seems to be more frequent in classroom lessons and broadcast

dialogues than in private dialogues. However, whenever *ah* is used in these two public text types, teachers/lecturers and guests in broadcast dialogues use *ah* exclusively in direct speech when imitating opinions of other people – very often views of the average citizen. In (15), a lecturer uses *ah* in direct speech when imitating common criticism of the statistical modelling she is presenting.

Eh is found across all four text types but there seems to be a slight preference for private dialogues. *Eh* also seems fairly frequent in legal cross-examinations but the four tokens of *eh* in this text type are very specific and rather unusual for these very formal interactions as all of them appear in rare antagonistic disputes between attorneys and witnesses (16) is an excerpt from a very heated controversy between an attorney and a witness, which eventually results in an intervention of the police in court who take violent action against the witness. The attorney uses *eh* to emphasize his verbal attack at the witness. *OK* is found most commonly in classroom lessons. Teachers frequently use *OK* in a facilitative way to check whether the students are still following their explanations, have understood everything, or have a question, as in (10) and (17). No *right* tokens were found for legal cross-examinations, but the distribution across the other three text types is fairly balanced. *You know* is used across all four text types but is less frequent in legal cross-examinations. The text
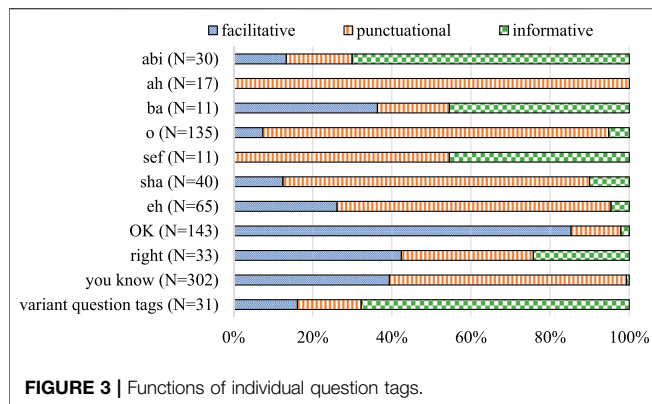
FIGURE 3 | Functions of individual question tags.

type distribution of variant question tags seems quite even according to the descriptive statistics.

15) <#>Anybody can say *ah* this test is not reliable (les_08)
16) <#>I will disgrace you more than that *eh* (cr_09)
17) <#>And one of the important concepts in discourse analysis is cohesion *OK* <#>Cohesion and Coherence *OK* (les_03)

In terms of the functions of question tags, punctuational uses dominate (653; 49.2%) followed by facilitative question tags (452; 34.1%). Only 138 question tags (10.4%) were classified as informative. 83 question tags were coded as multifunctional, which mostly serve a facilitative and punctuational function (74; 5.6%). Question tags that combine an informative with punctuational (5; 0.4%) or facilitative (4; 0.3%) functions are rare. While many question tag forms are used for all three functions, there are preferential differences as some forms have a more focused form-function mapping. **Figure 3** shows the functional diversity of the most frequent English and non-English question tag forms; multifunctional tags were excluded.

Ah has the most focused functional profile as it only serves punctuational functions. This functional exclusivity suggests that the classification of *ah* as a question tag is somewhat problematic as it seems that this form cannot fulfill all functions that are central to the definition of a question tag in this analysis. However, the findings for *ah*, similar to *ba* and *sef*, should be treated with some caution due to the low token frequency. *O* and *sha* also have a very focused form-function relationship with a strong preference for punctuational uses. *Abi* has a slightly more diverse functional profile but is preferentially used informatively. The functional diversity of *ba* is quite balanced but there is a slight dispreference for punctuational uses. No token of *sef* was classified as facilitative and there is an almost even distribution between punctuational and informative uses. *Eh* may fulfill all three functions but punctuational uses clearly dominate. *Right* has a functional profile that is quite balanced but there is a slight preference for facilitative uses. Both *OK* and *you know* have a rather focused functional profile as *OK* is mainly used facilitatively and *you know* punctuationally. For variant question tags the descriptive statistics demonstrate a preference for informative uses.

The descriptive statistics and examples have illustrated multilingual variation in question tag forms and the effects of text type on question tag use. In addition, this section has shown the form-function relationship of the most frequent English and non-English question tag forms in a descriptive way. As both text type and function influence which question tag forms Nigerian speakers select in dialogues, a further multivariate inferential analysis is necessary to highlight the details of variation. Further inferential statistics are not possible for question tag forms with a low token frequency or forms that exhibit (almost) categorical variation, i.e., *abi*, *ah*, *ba*, *eh*, *right*, *sef*, and *sha*.

The regression models demonstrate the effects of text type and function on the selection of *o*/*OK*/*you know*/variant question tags in contrast to all other question tags. These results are shown in **Tables 3–6**. All forms have a specific functional profile. For *o*, there is a strong preference for punctuational uses and the Nigerian speakers show a preference for using *OK* in a facilitative way. *You know* is preferentially used for punctuational and facilitative functions and there is a strong dispreference for informative uses. In contrast, variant question tags are preferentially selected for informative uses and dispreferred for both other functions.

Text type has a significant effect on the use of *o*, *OK*, and *you know*. However, the results for text type variation are somewhat vulnerable to the low token frequencies in legal cross-examinations. Consequently, these results need to be viewed with some caution despite the use of speaker as a random factor, which helps to prevent Type I errors. As already shown in the descriptive statistics, there is a strong preference for *o* in private dialogues, while it is dispreferred in classroom lessons and broadcast discussions. Legal cross-examinations demonstrate no tendency. For *OK*, there is a strong preference for classroom lessons and a somewhat lower preference for legal cross-examinations, while both other text types exhibit a dispreference. For *you know*, there is a strong preference in broadcast dialogues, no tendency in private dialogues, and a dispreference in classroom lessons and legal cross-examinations.

Hence for most question tags, text type and function together constrain their use. This combination of predictors reflects the communicative needs of speakers in specific situations. *O* is commonly used in informal conversations among friends to add emphasis to arguments, humorous remarks, or expressions of surprise. *O* is often used when speakers become emotionally involved in a discussion, as in (18) where a student expresses her anger against certain study regulations.

18) <#>*Ah ah* god will not allow it *o* (con_13)

OK is a typical teacher question tag used as a pedagogical strategy in longer explanatory passages to integrate the students into the discourse as already highlighted in (10) and (17). *You know* is used very frequently by hosts and guests in broadcast discussions to add emphasis to statements and to integrate the other participants into the discussion. In (19), the host talks about interethnic relationships in Nigeria and expresses her joy about a famous interethnic couple. She uses *you know* to emphasize her emotions and to invite the other participants to contribute to the discussion.

TABLE 3 | Regression analysis *o*; effects of text type and function.

$R^2 = 0.56$		N	%-o	(Centered) factor weight
text type; p <0.001	private dialogue	720	16.8	0.79
	legal cross-examination	44	4.5	0.50
	broadcast dialogue	219	3.7	0.36
	classroom lesson	260	1.5	0.33
function; p <0.001	punctuational	653	18.1	0.80
	facilitative	452	2.2	0.35
	informative	138	5.1	0.32

TABLE 4 | Regression analysis *OK*; effects of text type and function.

$R^2 = 0.67$		N	%-OK	(Centered) factor weight
text type; p <0.001	classroom lesson	260	32.7	0.82
	legal cross-examination	44	6.8	0.65
	broadcast dialogue	219	13.2	0.35
	private dialogue	720	3.6	0.18
function; p <0.001	facilitative	452	27.0	0.86
	punctuational	653	2.8	0.34
	informative	138	2.2	0.24

TABLE 5 | Regression analysis *you know*; effects of text type and function.

$R^2 = 0.70$		N	%-you know	(Centered) factor weight
text type; p <0.001	broadcast dialogue	219	47.5	0.81
	private dialogue	720	18.5	0.51
	classroom lesson	260	23.8	0.36
	legal cross-examination	44	6.8	0.28
function; p <0.001	punctuational	653	27.7	0.77
	facilitative	452	26.3	0.76
	informative	138	1.4	0.08

TABLE 6 | Regression analysis variant question tags; effects of function.

$R^2 = 0.61$		N	%-Variant question tag	(Centered) factor weight
function; p <0.001	informative	138	15.2	0.91
text type; p = 0.26	facilitative	452	1.1	0.26
	punctuational	653	0.8	0.21

19) <#>Sometimes when I see them on screen I could just see love radiating between the two <#>Two handsome and beautiful looking individuals *you know* (bdis_05)

Variant question tags tend to be used in an informative way across all four text types irrespective of whether the exact question tag form agrees to the main clause or not, as shown in (1, 2, 3, 13, and 14). Due to the overall low token frequency of variant question tags, a more detailed analysis of the exact form (e.g. variant vs. invariant uses or polarity) in relation to form and function was not possible.

While the formality of text types plays a crucial role for variation in question tag use, particularly for non-English forms, text type variation is more complex than an informal-

formal dichotomy suggests. For many question tags, the speakers' roles and the particular communicative needs that go along with that role in a given context are decisive for the selection of a particular question tag form.

RESULTS II: SURVEY STUDY OF THE PERCEPTION OF QUESTION TAGS AMONG NIGERIAN STUDENTS

The survey study illustrates which question tag forms Nigerian students prefer in different situations and which attitudes they hold toward a selection of English and non-English forms. Overall, the survey shows a strong preference among the

FIGURE 4 | Results of multiple-choice task (*N* = 49). Classroom lesson and legal-cross examination are abbreviated as classlesson and legalcrossex, respectively.
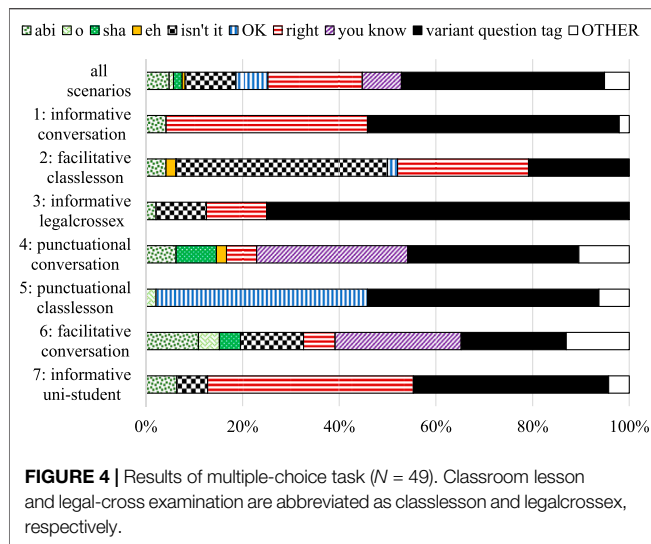
participants for variant question tags over invariant ones, which contrasts the corpus-based results on language use. Moreover, English forms are generally preferred over non-English ones in both parts of the survey. **Figure 4** demonstrates the participants' choices of question tag forms in the multiple-choice task in percent across all seven scenarios and for each one individually. Overall, variant question tags were selected most frequently (42.0%), followed by *right* (19.5%), invariant *isn't it* (10.5%), *you know* (8.1%), and *OK* (6.6%). *Eh* and all non-English forms were strongly dispreferred. Of these forms *abi* was selected the most frequently, being chosen in 4.7% of all cases.

The multiple-choice task also shows that speakers' choices for particular forms were constrained by the communicative settings and question tag functions targeted by the different scenarios. The preference for variant question tags is most pronounced in scenarios three, one, and five; the former two require the use of an informative question tag. *Right* was mostly chosen for informative uses in scenarios one and seven. At the same time, *right* was also chosen relatively frequently in scenario two.

Regression analyses that investigate how scenario affects the selection of one form in contrast to all others for variant question tags and *right* demonstrate that scenario has a significant effect on the selection of either *right* or variant question tags, corroborating the descriptive results. The results in **Table 7** illustrate that there is a strong preference for the selection of variant question tags in scenarios three, one, and five, while variant question tags are strongly dispreferred for facilitative uses in scenarios two and six. **Table 8** shows that for *right*, there is a strong preference for informative uses as well, as given in scenarios one and seven. In addition, *right* is also a relatively frequent form in scenario two, which is a classroom lesson and targets a facilitative question tag. For punctuational uses and facilitative uses in a conversation *right* is dispreferred. The level of formality of a situation seems to be decisive for the difference between the selection of variant question tags and *right* for informative uses as variant question tags clearly dominate over *right* in the scenario depicting a legal cross-examination.

The effect of the communicative setting and function are also evident for the other question tag forms that are overall less frequent and were hence not investigated via inferential statistics. For example, *OK* was only selected for scenario five, which depicts a classroom lesson and requires a punctuational question tag. *You know* was only selected for informal conversations between friends, either for punctuational (scenario 4) or facilitative (scenario 6) uses. Non-English question tags were only selected in scenarios depicting informal private dialogues. For example, 19% of students selected either *abi*, *sha*, or *o* in scenario 6. For invariant *isn't it*, no clear usage profile emerges from the multiple-choice task.

The Matched-Guise-Test also demonstrates a clear preference for variant question tags and English invariant ones, while non-English forms and *eh* are perceived less positively. **Figure 5** shows the mean ratings of the nine question tag forms (as different lines) on the six items on six-point scales. Values below 3.5 are negative and values above 3.5 positive. Generally, all speakers using question tags were rated rather positively, typical of an acquiescence bias (Garrett, 2010: 45). The descriptive statistics illustrate that the variant question tags seem to be rated overall most positively across the six items. In contrast, speakers using *abi*, *eh*, *o*, and *sha* were rated least positively. The ratings of invariant *isn't it*, *OK*, *you know*, and *right* fall in between these poles; with speakers using *right* being rated somewhat less positively than the other four question tags.

Figure 5 also demonstrates that the ratings of the individual question tags vary substantially according to the rating item, but this level of variation is difficult to interpret. To facilitate the further interpretation of the data a Principal Component Analysis was run to extract individual components. Due to the rather small sample size, Principal Component Analysis is not ideal and the outcome should not be interpreted in mathematical detail but shows the general underlying structure of variation in the sample.[7] Principal Component Analysis was forced to extract two components and Varimax with Kaiser Normalization was used as a rotation method. The Rotated Component Matrix was used to interpret which items form a cluster. According to this model, polite, educated and gentle clustered together to a first component, which has an eigenvalue of 4.4 and explains 75% of the variation. The factor loadings of the three items are very similar (polite: 0.88; educated: 0.87; gentle: 0.83). This first component was labelled 'decency'. The second component is less central to the variability in the data as it has an eigenvalue of 0.6 and only explains 11% of the variation. Outgoing, direct, and friendly loaded onto this second component. With a factor loading of 0.92, outgoing is most central to the second component, followed by direct (0.76). Friendly is least important to the second component with a factor loading of only 0.62, and friendly also loaded onto the first component (0.64). However, friendly was still grouped to the second component as friendly fits outgoing and direct better on a conceptual level. This second component was labeled

---

[7]See Field (2009: 636–650) for a detailed discussion of sample size, reliability, and validity of Principal Component Analysis.

TABLE 7 | Regression analysis on the effect of scenario on the selection of variant question tags in contrast to all other forms in the multiple-choice task (N = 49).

Scenario p <0.001; R^2 = 0.52	%-Variant question tag	(Centered) factor weight
3 (informative, legal cross-examination)	73.5	0.89
1 (informative; conversation)	51.0	0.66
5 (punctuational; classroom lesson)	46.9	0.60
7 (informative, uni-student)	38.8	0.48
4 (punctuational; conversation)	34.7	0.42
2 (facilitative; classroom lesson)	20.4	0.20
6 (facilitative; conversation)	20.4	0.20

TABLE 8 | Regression analysis on the effect of scenario on the selection of *right* in contrast to all other forms in the multiple-choice task (N = 49)tbl8fnlowast[a].

Scenario p <0.001; R^2 = 0.25	%-*right*	(Centered) factor weight
1 (informative; conversation)	40.8	0.76
7 (informative, uni-student)	40.8	0.76
2 (facilitative; classroom lesson)	26.5	0.63
3 (informative, legal cross-examination)	12.2	0.39
4 (punctuational; conversation)	6.1	0.23
6 (facilitative; conversation)	6.1	0.23
5 (punctuational; classroom lesson)	0.0	-

[a]*Scenario 5 was excluded from the analysis due to categorical absence of* right.
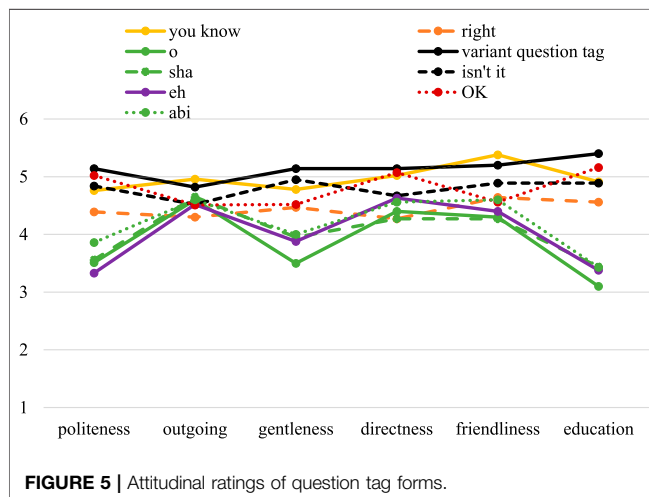
FIGURE 5 | Attitudinal ratings of question tag forms.

'expressiveness'. For further analysis, ratings for decency and expressiveness were calculated as mean values of the respective three items that loaded on the specific component.

Logistic regression modelling was used to further analyze variation in the decency and expressiveness ratings for the nine question tags. Decency and expressiveness ratings were used as the dependent variables in two separate models and question tag form as a fixed predictor variable. **Table 9** shows the mean values for the decency and expressiveness ratings as well as the results for the regression models. Variation for the decency ratings is much more pronounced than for expressiveness. For the former ratings, there is a range of 1.83, while for the latter the range is only 0.74. For both dimensions, variant question tags and

invariant English question tags were rated more positively than *eh* and the three non-English forms. For decency, these differences in the ratings reach the level of significance and the coefficients clearly illustrate this rating pattern. There are positive coefficients for the English forms except for *eh*, and negative coefficients for the non-English forms and *eh*. However, for expressiveness, the differences in the ratings did not reach the level of significance. Hence, no coefficients and R^2 are reported.

Open comments on the different question tags were rare in the surveys. Most comments were made for the non-English forms *abi*, *o*, and *sha*. In many comments, participants tried to explain the specific pragmatic functions of the forms, for example highlighting that "the use of o is showing emphasis", or gave English translations, such as "abi is a Nigerian slang for right". Opinions were very divided in terms of the 'correctness' of the use of non-English question tags. One student commented that "The use of 'abi' is really wrong in English language. That is code-mixing. Mixing Yoruba and English". In contrast, others expressed their appreciation of these question tags: "I love that use of 'o' like that". Several students commented that the use of *o*, *abi*, or *sha* does not indicate a low level of education but is a common form of code-mixing in Nigeria: "abi is a Yoruba word, if it is used it doesn't mean the user is uneducated. It's just code mixing". Students also stated that the use of non-English forms signals and also requires intimacy and a close relationship between interlocutors: "it is not proper to use abi for someone you just met". Several comments highlighted that students generally perceived the use of question tag forms as informal: "speaker is informal with the use of you know". The few comments made do not allow any closer analysis but illustrate that speakers are aware of the diversity of question tag forms and that these forms carry diverse indexicalities in Nigerian English.

DISCUSSION AND CONCLUSION: THE MULTILINGUAL PRAGMATICS OF NIGERIAN ENGLISH

This study has illustrated multilingual variation in question tag use in Nigerian English and Nigerian students' perception of this variation. The corpus-pragmatic analysis of six dialogue text types of ICE-Nigeria has shown that Nigerian speakers use a wide range of English and non-English question tag forms. Similar multilingual variation has been described for Cameroonian (Mbakop, 2020), Philippine (Westphal, 2020),

TABLE 9 | Decency and expressiveness ratings; regression analysis on the effect of form on ratings.

Decency p <0.001; intercept = 4.29; R^2 = 0.11				Expressiveness p = 0.373		
Form	Coef	N	Mean	Form	N	Mean
variant question tag	0.93	43	5.22	*you know*	48	5.13
isn't it	0.60	44	4.89	variant question tag	45	5.08
OK	0.60	44	4.89	*OK*	44	4.73
you know	0.53	45	4.82	*isn't it*	45	4.71
right	0.18	47	4.47	*eh*	46	4.58
abi	−0.50	43	3.79	*abi*	44	4.57
sha	−0.64	43	3.65	*o*	46	4.47
eh	−0.78	43	3.51	*right*	47	4.40
o	−0.90	47	3.39	*sha*	45	4.39

and Trinidadian English (Wilson et al., 2017). In contrast to the overwhelming focus on variant question tags in previous research (e.g., Tottie and Hoffmann, 2006; Borlongan, 2008; Barron, 2015; Parviainen, 2016; Hoffmann et al., 2017) in World Englishes, these types of question tags are marginal in Nigerian English as they only account for 2.5% of occurrences. Invariant uses make up 54.5% of these canonical question tag structures. However, they still cannot be viewed as particularly characteristic of Nigerian English in contrast to the many indigenous Nigerian forms, such as *abi*, *o*, or *sef*.

The strong dominance of invariant forms in general and the use of many indigenous forms is in line with other studies on New Englishes that have also investigated variant and invariant question tags (Wilson et al., 2017; Mbakop, 2020; Westphal, 2020). In contrast (Gómez González, 2018: 122) shows a higher frequency of variant than invariant question tag forms for British English. Taken together this means that New Englishes are not necessarily characterized by invariant uses of *isn't it* or *is it (not)* as argued by Borlongan (2008), Parviainen (2016), and Hoffmann et al. (2017) but by the usage of a wide range of English and particularly non-English invariant question tags. Hence, future research on question tags in New Englishes needs to operationalize question tags as a multilingual set of diverse forms and should not be restricted to 'canonical' ones.

The multilingualism of question tags in Nigerian English is also evident for the form *now*, which seems to combine nativized patterns of English *now* (Oladipupo and Unuabonah, 2021) and Nigerian Pidgin *na* (Unuabonah et al., 2021). Furthermore, the high frequency of *you know* (71.2tpf) in contrast to Cameroonian (52.8tpf; Mbakop, 2020), Philippine (26.6tpf; Westphal, 2020), and Trinidadian English (43.4tpf; Wilson et al., 2017) suggests that this form might also be characteristic of Nigerian English in contrast to other New Englishes, but *you know* requires a closer cross-variety analysis.

Considering the high frequency of non-English forms, the fluid integration of these forms into Nigerian English, and their alternation with English forms, the strict distinction into English and non-English seems inapt. All question tag forms identified in the analysis, whether *abi*, *right*, or variant question tags, are part of the repertoire of speakers sampled in ICE-Nigeria when speaking English. Hence, all forms are part of Nigerian English but just have different origins and different usage patterns.

Text type has a strong effect on the frequency of all forms taken together and individual ones. Question tags are overall most frequent in private dialogues, followed by classroom lessons, broadcast dialogues, and legal cross-examinations. Hence, formality and the degree of pre-structuring seem to play a decisive role. The more informal the text type and the less predefined the communicative situation, the higher the frequency of question tags. Hence, question tags can be viewed as somewhat informal structuring devices in dialogues. In addition, the particular communicative setting/conventions of a text type and how these are realized are decisive for the use of question tags. In classroom lessons, there is a teacher who leads the dialogue with the students and has to integrate them into the classroom discourse. However, the analysis has shown that there are different teaching styles in classroom lessons in ICE-Nigeria. Some teachers have a quasi-written teaching style, dictating texts to the students and quoting extensively from written material, and use very little question tags. In contrast, other teachers have a more interactive teaching style, speaking freely and trying to integrate the students into the discourse by means of question tags.

Similar to Wilson et al. (2017) and Westphal (2020), text type also has a strong effect on the selection of particular question tag forms over others. The analysis has shown that most non-English question tag forms are mainly used in informal private dialogues, while they are rare in more formal public text types. In addition to text type, function is decisive for the use of specific question tags. Several forms have a very focused functional profile. For example, *ah*, *eh*, and *o* are dominantly used in a punctuational way, while *abi* and variant question tags are preferentially used for informative functions. These two factors interact and together influence the choice of individual forms. For example, the choice between *abi* and variant question tags is controlled by the formality of the text type. *Abi* is restricted to informal situations and variant question tags are used across all text types. *OK* is mainly used facilitatively, and Nigerian teachers used it particularly often for this purpose, similar to Trinidadian and Philippine teachers (Wilson et al., 2017; Westphal, 2020). Hence, the choice of specific forms is determined by the communicative needs of speakers in accordance with their discursive role. For example, *you know* is very versatile in ICE-Nigeria being used in facilitative and punctuational ways. *You know* is especially frequent in broadcast discussions. Hosts mainly employ it facilitatively to integrate the audience and the

guests into the conversation. In contrast, guests use *you know* mostly in a punctuational way to emphasize their arguments.

The survey study contrasts the findings for language use strongly as it has shown a general preference for variant question tags over invariant forms and non-English forms are strongly dispreferred. Despite the general bias toward variant question tags, the students' choices in the multiple-choice task are guided by the particular communicative settings and demands targeted in the different scenarios. These constraints on the students' selections overlap with the usage profiles of individual question tags. For example, variant question tags were selected especially often for informative uses and *OK* for a classroom situation. *You know* was only selected for facilitative and punctuational uses in informal private dialogues. This means that the test – which is the first of its kind for discourse-pragmatic-features – works, as scenario has a significant effect on the students' choices.

The Matched-Guise-Test has shown that the participants' attitudes are guided by a prescriptive ideology (e.g., Osakwe, 2009; Mbakop, 2020: 1), which positions speakers using variant question tags as more polite, gentle, and educated. In contrast, *abi*, *o*, and *sef* were rated less positively in terms of the items reflecting the attitudinal dimension of decency. The strong dominance of this prescriptive ideology was further supported by comments devaluing non-English forms as incorrect for English. However, there were no differences in the ratings of the different question tags on items reflecting expressiveness. This means that all speakers using question tags were viewed as being more outgoing, direct, and friendly – irrespective of the particular form. The open comments also illustrated a certain pride in the Nigerian forms *abi*, *o*, and *sef*, which are viewed to mark a Nigerian identity when speaking English.

The mixed-methods approach of this study has illustrated the multilingual pragmatics of Nigerian English from two perspectives. On the one hand, indigenous Nigerian languages and Nigerian Pidgin have been shown to have a substantial influence on the use of question tags in dialogues from ICE-Nigeria. The Nigerian speakers sampled in the corpus combine English and non-English forms in fluid ways as forms originating from indigenous language are well-integrated into Nigerian English albeit to different degrees. For example, *o* is used frequently across a wide range of communicative situations and can be used for different pragmatic purposes. In contrast, several indigenous forms, such as *ba* and *sef*, are rarer and restricted to informal conversations, and have much narrower pragmatic profile, such as *ah*. As already argued, *now* is especially characteristic for the multilingual pragmatics of Nigerian English but requires further socio-pragmatic and socio-phonetic analysis. Both text type and function have a significant effect on the multilingual variation of question tag use and indicate that sociolinguistic factors should be considered when describing the multilingual dynamics of New Englishes, which exhibit substantial internal variation. Future variational pragmatic research on the multilingual pragmatics of Nigerian English or other New Englishes may also take into account macro-social factors, such as age and gender, but should still pay close attention to the particular roles speakers have in the contexts under analysis.

On the other hand, the survey study has shown that the acceptance of indigenous and Nigerian Pidgin forms is clearly lagging behind their widespread use in English. This difference between corpus and survey data for pragmatic phenomena is reminiscent of Schneider K. P. (2007) analysis of the speech act of thanking responses. Both perspectives are important for the assessment of the degree of nativization of Nigerian English (Schneider E., 2007). In terms of multilingual variation for question tags, the study suggests that there is a high degree of nativization in Nigerian English at this pragmatic level of variation, but there is a strong 'complaint tradition' (Schneider E., 2007: 43), which devalues local innovations. Further qualitative interviews (e.g., Anchimbe, 2018) are needed to investigate Nigerians' perspective on multilingual variation in question tag use in more detail.

On a purely methodological level, this study has worked with a very wide understanding of what counts as a question tag relying on function-based criteria in contrast to much previous research that has used formal criteria (e.g., Tottie and Hoffmann, 2006; Borlongan, 2008; Barron, 2015; Parviainen, 2016; Hoffmann et al., 2017). The current approach is not unproblematic as many forms are included in the analysis that have a very focused functional profile and are rarely used for informative purposes. For example, *you know* and *o* are mainly used punctuationally, *ah* is exclusively used in this way, and *OK* mostly serves facilitative functions. In addition, the participants did not select *you know*, *o*, and *OK* for informative uses in the multiple-choice-test. Hence there is a dispreference for informative uses and the pragmatic profile of these forms differs significantly from variant question tags, which are mainly used informatively by the Nigerian speakers and selected most often for informative uses in the survey. However, variant question tags also fulfill other functions and are also selected by the Nigerian students for punctuational and facilitative uses in the survey. In addition, previous research on variant question tags in other varieties has demonstrated that while informative uses are very frequent, the majority of question tags is used in other ways. For example, informative (i.e., informative and confirmatory uses combined) only account for 36.9% in Tottie and Hoffmann (2006: 302) or for 32.6% in Borlongan (2008: 14). This means that there is still enough pragmatic overlap of *ah*, *o*, *OK*, *you know* with variant question tags to conceptualize all these forms as question tags. In conclusion, a function-based conceptualization of question tags casts a much wider net and shows a broader range of variation. This approach may be problematic as one of the defining criteria for sociolinguistic variables of equivalence of meaning is violated, as is generally the case for using discourse-pragmatic-features for a variationist analyses (Pichler, 2010; Pichler, 2013: 28–32). However, relying on specific formal criteria excludes many forms, which in the case of Nigerian English are integral parts of the speakers' linguistic repertoire. This choice between form and function reflects the tension in the field of corpus-pragmatics, which brings together the two very different disciplines of pragmatics, which is mainly concerned with pragmatic functions, and corpus-linguistics, which relies mostly on forms (see Aijmer and Rühlemann, 2015: 1–9).

On a further methodological level, the corpus-pragmatic analysis of ICE-Nigeria has shown that the ICE corpora may be suitable to analyze multilingual variation in New Englishes. However, it is essential to investigate multilingual variables. Discourse-pragmatic-features seem especially suitable for such an endeavor as speakers in New Englishes have been shown to integrate indigenous forms when speaking English (e.g., Unuabonah and Oladipupo, 2018; Unuabonah, 2019; Unuabonah, 2020; Westphal, 2020). However, in order to describe the sociolinguistic dynamics of multilingualism in New Englishes, analyses must consider English and indigenous forms. Excluding entire groups of variants may veil essential aspects of pragmatic variation and may lead to biased conclusions. Corpus-pragmatic analyses of entire sets of discourse-pragmatic-features in New Englishes should use pre-defined concordance lists that are based on previous research on British or American English with caution and should consider other forms or ways of expressing similar pragmatic functions targeted in the analysis. Such an approach may complicate cross-variety comparisons but may show further differences between Englishes spoken as a native language and New Englishes not covered when multilingual variants are excluded. Besides discourse-pragmatic-features, ICE-Nigeria may also be used for close qualitative analyses of code-switching as the corpus itself contains many instances of code-switching to indigenous Nigerian languages, which however are often not transcribed but are still accessible through the accompanying sound files. Such a detailed qualitative approach may also look more closely at the exact pragmatic functions of question tags in a given situation, which were operationalized in a very generalizing way in this quantitative study to allow regression analyses.

Finally, the corpus analysis of question tags has shown that text type variation is essential for describing the dynamics of multilingual variation. Standard New Englishes are by no means homogenous entities, but local innovations and indigenous languages are integrated to different degrees, which can be operationalized along different text types. Although the ICE corpora may be used to illustrate multilingual pragmatic variation in New Englishes, their possibilities are somewhat limited due to the corpora's main focus on (monolingual) Standard English. Hence, Mair's (2011: 234) call for the compilation of multilingual corpora for New Englishes still applies, as such corpora may be more suitable to illustrate the multilingual

pragmatics of New Englishes as well as multilingual variation on other levels of linguistic variation. Studies on New Englishes that describe multilingual variation are essential for a better understanding of their linguistic structure and hence for issues of description and codification. In addition to such studies on multilingual language use, research on New Englishes must also take into account the users' perspective on linguistic variation more earnestly, which may well differ from the findings on language use and provide additional insights not anticipated.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

K. Aijmer and C. Rühlemann (Editors) (2015). *Corpus Pragmatics* (Cambridge: Cambridge University Press).

Aijmer, K. (2013). *Understanding Pragmatic Markers in English*. Edinburgh: Edinburgh University Press.

Algeo, J. (1990). "It's a Myth Innit?: Politeness and the English Tag Question," in *The State of the Language*. Editors C. Ricks and L Michaels (Berkeley, CA: University of California Press), 443–450.

Anchimbe, E. (2018). *Offers and Refusals: A Postcolonial Pragmatics Perspective on World Englishes*. Amsterdam: Benjamins.

Anderson, J. A., and Ansah, G. N. (2015). "A Sociolinguistic Mosaic of West Africa: Challenges and Prospects," in *Globalising Sociolinguistics: Challenging and*

Expanding Theory. Editors D. Smakman and P. Heinrich (London: Routledge), 54–65.

Axelsson, K. (2011). A Cross-Linguistic Study of Grammatically-Dependent Question Tags. *Stud. Lang.* 35 (4), 793–851. doi:10.1075/sl.35.4.02axe

Barron, A. (2015). ""And Your Wedding Is the Twenty-Second of June Is it?"," in *Pragmatic Markers in Irish English*. Editor C. P. Amador-Moreno (Amsterdam: Benjamins), 203–228. doi:10.1075/pbns.258.09bar

Bautista, L. (2011). "Some Notes on 'no in Philippine English," in *Exploring the Philippine Component of the International Corpus of English*. Editor L. Bautista (Manila: Anvil), 75–89.

Beeching, K. (2016). *Pragmatic Markers in British English*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Borlongan, A. (2008). Tag Questions in Philippine English. *Philippine J. Linguistics* 39 (1), 1–34.

Columbus, G. (2009). "A Corpus-Based Analysis of Invariant Tags in Five Varieties of English," in *Corpus Linguistics: Refinements and Reassessments*. Editors A. Renouf and A. Kehoe (Amsterdam: Rodopi), 401–414.

Columbus, G. (2010). A Comparative Analysis of Invariant Tags in Three Varieties of English. *English World-Wide* 31 (3), 288–310. doi:10.1075/eww.31.3.03col

Deuber, D. (2005). *Nigerian Pidgin in Lagos: Language Contact, Variation and Change in an African Urban Setting*. London: Battlebridge.

Field, A. (2009). *Discovering Statistics Using SPSS*. Los Angeles: SAGE.

Garrett, P. (2010). *Attitudes to Language*. Cambridge: Cambridge University Press.

Gómez González, M. de los Ángeles (2018). "Variable and Invariable Tag Questions in Spoken British English," in *The Construction of Discourse as Verbal Interaction*. Editors M. de los Ángeles Gómez González and J. L. Mackenzie (Amsterdam: Benjamins), 109–144.

Greenbaum, S., and Nelson, G. (1996). The International Corpus of English (ICE) Project. *World Englishes* 15 (1), 3–15. doi:10.1111/j.1467-971x.1996.tb00088.x

Gut, U., and Fuchs, R. (2013). Progressive Aspect in Nigerian English. *J. English Linguistics* 41 (3), 243–267. doi:10.1177/0075424213492799

Gut, U. (2012). Towards a Codification of Nigerian English – the ICE Nigeria Project. *J. Nigeria English Stud. Assoc.* 15 (1), 1–13.

Hansen, B. (2018). *Corpus Linguistics and Sociolinguistics: A Study of Variation and Change in the Modal Systems of World Englishes*. London: Brill.

Hoffmann, S., Blass, A.-K., and Mukherjee, J. (2017). "Canonical Tag Questions in Asian Englishes," in *The Oxford Handbook of World Englishes*. Editors M. Filppula, J Klemola, and D Sharma (Oxford: Oxford University Press), 697–714.

Hundt, M., and Gut, U. (Editors) (2012). *Mapping Unity and Diversity World-wide: Corpus-Based Studies of New Englishes* (Amsterdam: Benjamins).

Johnson, D. E. (2009). Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Lang. Linguistics Compass* 3 (1), 359–383. doi:10.1111/j.1749-818x.2008.00108.x

Jowitt, D. (2019). *Nigerian English*. Berlin: de Gruyter.

Kachru, B. B. (1985). "Standards, Codification and Sociolinguistic Realism: the English Language in the Outer circle," in *English in the World: Teaching and Learning the Language and Literatures*. Editors R. Quirk and H.G. Widdowson (Cambridge: Cambridge University Press), 11–30.

Kimps, D. (2018). *Tag Questions in Conversations: A Typology of Their Interactional and Stance Meanings*. Amsterdam: Benjamins.

Kortmann, B., and Schneider, E. (Editors) (2008). *A Handbook of Varieties of English* (Berlin: de Gruyter).

Leimgruber, J. (2013). *Singapore English: Structure, Variation, and Usage*. Cambridge: Cambridge University Press.

Mair, C. (2011). "Corpora and New Englishes: Using the 'Corpus of Cyber-Jamaican' to Explore Research Perspectives for the Future," in *A Taste for Corpora*. Editors F. Meunier, G Gilquin, and M Paquot (Amsterdam: Benjamins), 209–236. doi:10.1075/scl.45.13mai

Mbakop, A. W. N. (2020). Question Tags in Cameroon English. *English Today*, 1–11. doi:10.1017/s0266078420000139

Mesthrie, R. (2008). "Introduction," in *Varieties of English Vol 4: Africa, South and Southeast Asia*. Editor R. Mesthrie (Berlin: de Gruyter), 23–31.

Mufwene, S. S. (1994). New Englishes and Criteria for Naming Them. *World Englishes* 13 (1), 21–31. doi:10.1111/j.1467-971x.1994.tb00280.x

Oladipupo, R. O., and Unuabonah, F. O. (2021). Extended Discourse-Pragmatic Usage of Now in Nigerian English. *World Englishes* 40 (3), 371–389. doi:10.1111/weng.12492

Osakwe, M. (2009). The Tag Question Problem in Nigerian English Usage: Methodological Steps Out of the Gamble for Acceptability. *Lagos Notes Rec.* 15 (1), 18–29. doi:10.4314/lnr.v15i1.46449

Oyebola, F., Ho, B. S. Y., and HoLi, Z. (2019). "A Sociophonetic Study on the Variation in Educated Nigeran English," in Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia, August 2019. Editors S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Canberra, Australia: Australasian Speech Science and Technology Association Inc.), 2320–2323.

Parviainen, H. (2016). The Invariant Tagisn't Itin Asian Englishes. *World Englishes* 35 (1), 98–117. doi:10.1111/weng.12168

Pichler, H. (2010). Methods in Discourse Variation Analysis: Reflections on the Way Forward. *J. Sociolinguistics* 14 (5), 581–608. doi:10.1111/j.1467-9841.2010.00455.x

Pichler, H. (2013). *The Structure of Discourse-Pragmatic Variation*. Amsterdam: Benjamins.

Platt, J., Weber, H., and Ho, M. L. (1984). *The New Englishes*. London: Routledge.

Schneider, K., and Barron, A. (Editors) (2008). *Variational Pragmatics* (Amsterdam: Benjamins).

Schneider, E. (2007). *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press.

Schneider, K. P. (2007). Genre Matters: Textual and Contextual Constraints on Contemporary English Speech Behaviour. *Anglia* 125 (1), 59–83. doi:10.1515/angl.2007.59

Schneider, K. P. (2012). "Pragmatics," in *Areal Features of the Anglophone World*. Editor R. Hickey (Berlin: de Gruyter), 463–486. doi:10.1515/9783110279429.463

Schneider, E. W. (2016). Hybrid Englishes: an Exploratory Survey. *World Englishes* 35 (3), 339–354. doi:10.1111/weng.12204

Schröder, A., and Schneider, K. (2018). Variational Pragmatics: Responses to Thanks, and the Specificity of Namibian English. *English World-Wide* 39 (3), 338–364. doi:10.1075/eww.00017.sch

Siemund, P. (2013). *Varieties of English: A Typological Approach*. Cambridge: Cambridge University Press.

Takahashi, M. (2014). A Comparative Study of Tag Questions in Four Asian Englishes from a Corpus-Based Approach. *Asian Englishes* 16 (2), 101–124. doi:10.1080/13488678.2014.908529

Tottie, G., and Hoffmann, S. (2006). Tag Questions in British and American English. *J. English Linguistics* 34 (4), 283–311. doi:10.1177/0075424206294369

Unuabonah, F. O., and Gut, U. (2018). Commentary Pragmatic Markers in Nigerian English. *English World-Wide* 39 (2), 190–213. doi:10.1075/eww.00010.unu

Unuabonah, F. O., and Oladipupo, R. O. (2018). "You're Not Staying in Island Sha O": O, Sha and Abi as Pragmatic Markers in Nigerian English. *J. Pragmatics* 135, 8–23. doi:10.1016/j.pragma.2018.07.007

Unuabonah, F. O., and Oladipupo, R. O. (2021). Bilingual Pragmatic Markers in Nigerian English. *World Englishes* 40 (3), 390–406. doi:10.1111/weng.12453

Unuabonah, F. O., Oyebola, F., and Gut, U. (2021). "Abeg Na! We Write So Our Comments Can Be Posted!": Borrowed Nigerian Pidgin Pragmatic Markers in Nigerian English. *Pragmatics* 31 (3), 455–481. doi:10.1075/prag.19038.unu

Unuabonah, F. O. (2019). Frequency and Stylistic Variability of Discourse Markers in Nigerian English. *Corpus Pragmatics* 3 (3), 249–271. doi:10.1007/s41701-019-00059-8

Unuabonah, F. O. (2020). "So You Know Ehn . . . " The Use of Bilingual Interjections in Nigerian English. *Intercultural Pragmatics* 17 (2), 151–176. doi:10.1515/ip-2020-0008

Westphal, M. (2020). Question Tags in Philippine English. *Corpus Pragmatics* 4 (1), 401–422. doi:10.1007/s41701-020-00078-w

Wilson, G., Westphal, M., Hartmann, J., and Deuber, D. (2017). The Use of Question Tags in Different Text Types of Trinidadian English. *World Englishes* 36 (4), 726–743. doi:10.1111/weng.12247

Wunder, E.-M., Voormann, H., and Gut, U. (2010). The ICE Nigeria Corpus Project: Creating an Open, Rich and Accurate Corpus. *ICAME J.* 34, 78–88.

Investigating English in Multilingual Contexts Online: Identity Construction in Geotagged Instagram Data

Susanne Mohr*

Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway

Tourism discourse, referring to communication in tourism as global industry, contributes to the creation of tourist spaces, where space is a social and affective construct as opposed to place as a geographical one. Tourists and hosts are part of these spaces and form them with their language practices, both offline and online. This article presents a case study of tourism discourse related to Zanzibar on Instagram, focusing particularly on linguistic repertoires, the role of English in them and language choices as well as their implications for identity construction. A central issue, in line with discourse-centred online ethnography, is the comparison of the digital data with previously collected data from the physical tourist space. Theoretically and methodologically, the concept of (linguistic) transnationalism is central for the study, which uses geotags and hashtags as means of data retrieval and framework of analysis to further this concept.

Keywords: geotagging, hashtags, instagram, multilingualism, world englishes, zanzibar

INTRODUCTION

Increasingly large amounts of language are mediated digitally, and linguistic studies on computer-mediated communication abound (e.g., Herring, 1996; Thurlow and Mroczek, 2011; Tannen and Trester, 2013). Many of these focus on language practices specific to the digital sphere and social media, such as hashtagging (Zappavigna, 2015). The importance of language use online has also been acknowledged for the study of New and World Englishes (e.g., Heyd, 2016; Mair, 2018; Shakir and Deuber, 2019). However, many of these studies are based on large text corpora, comprising blog and website data, and do not analyze the specificities of microblogging on social media, whose importance has been emphasized by recent studies employing geotagged data from urban centers in the Global North (Androutsopoulos, 2014; Hiippala et al., 2019; Lyons, 2019). Investigations of geotagged social media data from less urban locations in the Global South remain scarce, however.

This article presents a case study of tourism discourse relating to such a space, i.e., Zanzibar, a semi-autonomous archipelago in the Indian Ocean, on Instagram. The social network is particularly relevant in this regard, given its focus on sharing (holiday) pictures, which are an important "souvenir" among tourists. The investigation, applying a discourse-centered online ethnographic approach (Androutsopoulos, 2008), focuses on the use of hashtags and their pragmatic functions. Some of these functions are similar to those of speech acts like greetings, with a strong focus on phatic communion, which has been proven to be central in tourism discourse (Mohr, 2020). The data stem from a corpus of geotagged posts by tourists and hosts, retrieved using a dedicated application programming interface (API). This method of data retrieval is particularly interesting for the study of World Englishes as it provides a very recent angle of analysis: while collected "in" a specific *place*

(Zanzibar), the posts offer a window of analysis for English(es) used to create *space* not tied to specific nation states, given the international background of tourists and hosts using Instagram, and away from traditionally studied academic contexts.

Correlating data from the digital tourist space with data and results on language choices and use in the physical tourist space is in line with the demand for relating offline and online data (Androutsopoulos 2008; Leppänen et al., 2020). The analysis also draws on previous *in situ* fieldwork in Zanzibar, which inspired the research focus on language choices and identity construction *via* language in the era of globalization. The central aims of this study are hence to find out 1) whether the digital tourist space is equally linguistically diverse as the physical tourist space, i.e., whether English is a "multilingua franca" (Jenkins, 2015), 2) whether there are specific linguistic patterns to be made out in hashtags that might be uncommon in the physical tourist space and 3) what the implications of the (lack of) linguistic diversity for the construction of identities in the digital tourist space of Zanzibar are.

To this end, the paper discusses Englishes and tourism discourse online in *Englishes and Tourism Discourse Online*, especially touching upon central issues concerning mobility, transnationalism, identity construction (online) and the Zanzibari tourist space. *Methods and Data* outlines methodological issues relating to data collection, coding and hashtags and their functions on social media, used here as an analytical framework. The results of the analysis are presented in *Results and Discussion*, touching upon linguistic patterns and themes observed in relation to identity construction. The final section provides some conclusions and an outlook on possible future research.

ENGLISHES AND TOURISM DISCOURSE ONLINE

World Englishes research, i.e., the study of features of varieties of English around the world as well as the socio-political and ideological implications of these varieties co-existing with other languages (Onysko, 2021), has recently been claimed to have reached an impasse (Saraceni, 2015; Saraceni and Jacob, 2021). The field seems not entirely up to date with recent developments in sociolinguistics and has only started to tackle issues such as globalization, superdiversity and digitization. More research in this vein is hence needed, also accounting more strongly for concerns such as multilingualism and translingualism. Among others, questions as to how English functions in international communication (i.e., considering the interface of World Englishes and English as a lingua franca (ELF) research) and how English is shaped by different modes and media will be central in future research (Onysko, 2021). While some literature on these issues exists (Friedrich and Diniz de Figueiredo, 2016; Bolander, 2020; Lim, 2020), more research inspired by a sociolinguistics of globalization framework (Blommaert, 2010) is desirable and it is one of the aims of this paper to provide such a study.

All these issues are closely related to Englishes being used outside of the traditionally studied academic circles in World Englishes, i.e., "at the grassroots" among non-elite speakers and in heterogeneous contexts with regards to acquisition and proficiency in English (Kubota, 2018; Meierkord and Schneider, 2021: 8). The present paper can contribute in this vein, given its focus on tourism discourse online.

(World) Englishes, Mobility and Transnationalism

The concepts of mobility, tourism, transnationalism, and globalization are closely linked. Transnationalism refers to "multiple ties and interactions linking people or institutions across the borders of nation states" (Vertovec, 1999: 447) and is probably most closely linked to globalization as "complex connectivity [. . .] the rapidly developing and ever-densening network of interconnections and interdependencies that characterize modern social life" (Tomlinson, 1999: 2). Bolander (2020: 678) comments on the difficulty of prying these two concepts apart, as they partly overlap. The differences between them depend on the conceptualization of both terms, however, and there is hence no definite answer. Some researchers have argued that transnationalism has a more limited purview (see also Bolander, 2020). Given that transnationalism is quite important for this article, it is discussed in more detail here.

Vertovec (1999) discusses six ways in which transnationalism has been conceptualized in research across different disciplines. Not all of them can be discussed or are indeed relevant here; I focus on the most pertinent aspects in the following. The most wide-spread notion is that of transnationalism as *social morphology*, i.e., a "social formation spanning borders" (Vertovec, 1999: 449), which is closely linked to migration and often applied in sociolinguistics (see Bolander, 2020). Tourism as the largest peaceful movement of people across cultural boundaries (Lett, 1989) can be conceptualized as a form of migration, or mobility at least (see Mohr, 2020). Thus, as people move across borders, new networks and communities of practice are formed, which constitute intriguing sociolinguistic environments worth studying. The notion of "community of practice" is usually fitting in these contexts offline but might be more complex in online spaces or at the nexus of both (King, 2019). This is elaborated on later, in *Constructing Identities Online*. Another conceptualization important for this paper is that of transnationalism as an *opportunity for capital*, in sociolinguistics often linked to commodified language practices in late capitalism (Bolander, 2020). Given that tourism is "the single largest international trade in the world" (Thurlow and Jaworski, 2017: 187), this is a central aspect in the study of the sociolinguistics of tourism and English in tourist contexts. This has been shown to be directly relevant for the physical tourist space of Zanzibar, where language use of English and the local language Kiswahili, as well as language learning are strongly commodified (Mohr, 2020; Mohr, 2021). Commodification and opportunities for capital are hence expected to play an equally central role for tourism discourse, i.e., "language and communication in tourism as global cultural industry" (Thurlow

and Jaworski, 2011: 222), online. Finally, transnationalism can be conceived as *hybrid cultural phenomena*, i.e., "facets of culture and identity [that] are [...] self-consciously selected, syncretized and elaborated from more than one heritage" (Vertovec, 1999: 451) and in that sense closely connected to transnationalism as the (*re*)*construction of place or locality*. Set in tourist spaces, the investigation presented in this paper focuses on how discourse and identity construction contribute to the construction of space, i.e., Zanzibar as a tourist destination, digitally.

Due to globalization, languages are generally less associated with nation states and "third wave" sociolinguistics focuses on speaker styles and globalized vernaculars as indexical of identities (Eckert, 2012). In World Englishes research, Bolander (2020: 682) argues, we need more investigations into "how varieties [...] are used as resources for constructing place and space, particularly in [...] territories where marking difference is/was coupled with drawing territorial distinctions". This is closely linked to the notion of *translingualism* as employed by Canagarajah (2020: 560) for instance, which is an extension of the concept of the language repertoire, moving beyond labelled national languages. The language practices in these repertoires are then utilized to construct space, a social and affective construct as opposed to place, a primordial geographical construct (Canagarajah, 2020: 559). In the study of transnational and translingual practices, it is important to document small scale contexts and their entanglement with global dynamics (Al Zidjali, 2019). This is what the current study aims at by documenting a relatively small space, i.e., Zanzibar, and its entanglement with the global dynamics of tourism.

A move away from nationalism and towards transnationalism does not only apply on the conceptual but also on the methodological level in World Englishes. An enhanced focus on linguistics of contact and language practices would be conducive to this (Bolander, 2020) and, in line with previous research on the physical tourist space of Zanzibar (Mohr, 2019; Mohr, 2020; Mohr, 2021), this is one of the present study's objectives as well. In this regard, the notion of the language repertoire (Benor, 2010) comprising various linguistic resources that a speaker can draw on depending on the context has been central and is again emphasized for this study.

The Physical Zanzibari Tourist Space

In order to further demonstrate some of the issues raised in the previous section and to contextualize the analysis provided later in this paper, a brief description of the physical Zanzibari tourist space is necessary. Zanzibar is an archipelago off the coast of Tanzania, of which it forms part as a semi-autonomous region. Thus, it is subject to the language policies of Tanzania, with Kiswahili as the *de facto* national language, spoken by ca. 47 million first and second language speakers, besides the 125 other languages spoken in the country. Kiswahili is also the lingua franca of Eastern Africa, with a considerable number of speakers in neighboring Kenya, some users in Burundi, Uganda and a few in Rwanda (Eberhard et al., 2021). Unguja island, the main island of Zanzibar, is also home to Kiunguja, the standard variety of Kiswahili, and the archipelago is hence an important linguistic center in East Africa. English is the *de facto* national working

language of Tanzania (Eberhard et al., 2021). In World Englishes research, the country has been argued to be an "Outer Circle" country where English is used as a second language (Schmied, 2008). However, it has been argued recently that English is in fact an additional language in Tanzania (for a recent account see Mohr, 2022). This seems different in Zanzibar, where tourism accounts for a major part of the local economy and many Zanzibaris learn English in order to be eligible for employment in the tourist sector (Mohr, 2020; Mohr, 2021).

Tourism has an important impact on language use on the archipelago. Thus, Zanzibaris spend a lot of financial resources on acquiring English semi-privately, i.e., in tuition classes taught by (retired) school teachers or in language classes at NGOs, given that many Zanzibaris do not think that language instruction in schools is sufficient to acquire the language fluently (see, Mohr, 2021). Besides English, many acquire additional foreign European languages like Italian or German in order to be even more attractive for the job market and to better accommodate tourists and their mother tongues (Mohr, 2019; Mohr, 2020). This closely relates to the abovementioned conceptualization of transnationalism as an opportunity for capital.

The most interesting phenomenon observed in this space and related to transnationalism as an opportunity for capital is what has been termed "Hakuna Matata Swahili" (HMS) (Nassenstein, 2019). It refers to a pidginized version of Kiswahili, allegedly developed for the tourists who usually do not speak the language (Mohr, 2020). It exhibits similarities to "mock languages", such as Mock Spanish as described by Hill (1998) for the United States. Interestingly, the Zanzibari participants in Mohr (2019, 2020) report that it is also frequently employed for communication not only by but with tourists as it supports selling touristic goods and services. To this end, it is also printed on souvenirs, postcards etc., thus making language directly purchasable.

In view of the many languages that are spoken in Tanzania, Zanzibar is a highly multilingual space and Zanzibaris have highly diverse language repertoires.[1] They draw on the various language practices comprised in these resources depending on different contexts and interlocutors (Mohr, 2019; Mohr, 2020). This is an illustrative example of translingual practices (see Canagarajah, 2020). When adding the tourists' various native languages and English to this already great linguistic diversity, a "superdiverse" space (Vertovec, 2007) is created that lends itself to investigations into multilingualism and language contact. In these spaces, English can hence be considered a "multilingua franca", i.e., one of many possible language choices in a multilingual space (Jenkins, 2015). This superdiversity is similar to digital spaces, which have been claimed to be inherently superdiverse, irrespective of tourism (see Leppänen et al., 2020). An investigation into a digital tourist space hence promises to be doubly intriguing.

---

[1]While there are fewer languages indigenously spoken in Zanzibar as compared to mainland Tanzania, especially work migration from the mainland and other parts of East Africa increases the linguistic diversity on the archipelago (Mohr, 2020).

Constructing Identities Online

Contemporary, "third wave" sociolinguistics (Eckert, 2012) is concerned with the indexicality of speech styles with regard to speaker identities. Speakers draw on various means of constructing meaning, i.e., verbal, discursive and semiotic resources, thus indexing (lack of) commonality, connectedness and groupness (Leppänen et al., 2020: 25). Nowadays, identities are increasingly viewed as temporary interactional positions "that social actors briefly occupy and then abandon as they respond to the contingencies of unfolding discourse" (Bucholtz and Hall, 2005: 591). Thus, identities are socially and situationally constructed and anything but fixed. In our globalized world, they are also complex and usually multiple, depending on individuals and groups alike. Turkle (1995: 180) mentions that "people experience identity as a set of roles that can be mixed and matched". These dynamics are closely linked to the concept of *community of practice* and related notions as discussed by King (2019). Online, where there might be no shared (strong) feeling of belonging among members of a group sharing similar interests, we might rather be concerned with affinity spaces (see also Gee, 2005), as more generalized social spaces (King, 2019). Publics, i.e., indefinite, unaccountable audiences addressed through circulated discourses or identity performances (Warner, 2002), are especially relevant with regard to social media. These concepts are often interrelated and co-occurring online, with affinity spaces containing sub-groups in the form of communities of practice, for instance (King, 2019). As will be shown in the analysis, the digital Zanzibari tourist space is also complex in this regard.

Situational dependency and complexity of identities specifically hold true for online digital spaces, where sociolinguistic research into identity work is a key topic (see Thurlow and Mroczek, 2011; Tannen and Trester, 2013; Leppänen et al., 2020). There, two factors that have also been shown to be central for the sociolinguistics of tourism in general and of Zanzibar in particular, are the *playfulness* of identity construction and the *intersection of identity work and authenticity* (Thurlow and Mroczek, 2011; Leppänen et al., 2015). Playfulness refers to language use being playful in tourist contexts in order to appeal to tourists on vacation, such as HMS in Zanzibar mentioned in the previous section. Authenticity, often aimed at by stylizing the location using a local language (see Salazar, 2006 on Tanzania), is another central aspect of many tourist spaces. With regard to digital spaces, playful language use has been attested frequently online and on social media, where users draw on different semiotic resources, such as emojis, pictures and verbal language, when communicating (e.g., Thurlow and Mroczek, 2011; Thurlow and Jaroski, 2020). Authenticity plays a role for the performance of identities online in that the disembodied nature of online communication apparently makes it easier for users to express themselves in ways that would normally be perceived as inauthentic (Turkle, 1995), for example using a variety of English spoken in another country. Chong (2020) reports an example from Singapore, where young Singaporeans employ different sociolinguistic resources to construct various online personae.

As the similarities between physical tourist spaces and online spaces demonstrate, there are parallels between offline and online language use. Thus, it has been argued that we do need research into connections between offline and online discourse, as these are an important aspect of digital language use (e.g., Androutsopoulos, 2008; Leppänen et al., 2020), especially where translocally connected groups are concerned (Blommaert, 2017). This would also thwart the too narrow focus on the particularities of digital discourse as present in the beginning of the field (Mair, 2020). With regard to identity work, it has been argued that online identities are an extension of offline identities, where language choices and code-switching illustrate different facets of a person's identities, depending on their perceived or desired social positioning (Barton and Lee, 2013: 55–68). This is particularly interesting in superdiverse spaces like the tourist space of Zanzibar (Mohr, 2020) or online spaces in general. It is this interface that the present paper aims at investigating.

METHODS AND DATA

Many studies into World Englishes online are based on large text corpora and do not analyze the specificities of microblogging and multimodal meaning making on social media in detail. Some studies do exist however, sometimes also employing geotagged data (e.g., Hiippala et al., 2019; Lyons 2019). Investigations of less urbanized locations in the Global South remain scarce, however. The data presented here stem from such a space, i.e., from tourism discourse associated with Zanzibar in the social network Instagram. The network's focus on pictures, which are a popular "souvenir" among tourists and often used to advertise services, hotels etc., among hosts, makes Instagram an ideal site for data collection. The general methodological framework applied is discourse-centered online ethnography (Androutsopoulos, 2008), combining participant observation of several Zanzibar tourism related accounts *via* an own Instagram account and the retrieval of posts via a dedicated API. The latter is described in more detail in the following section. In the future, interviews with key stakeholders are planned.

Data Collection

The move away from methodological nationalism (Schneider, 2019) and towards a more language contact oriented, transnational methodology mentioned earlier in *(World) Englishes, Mobility and Transnationalism* is one of the challenges this paper aims at addressing. As mentioned above, the investigation is based on Instagram posts retrieved using geotagged data from Zanzibar, that means Stone Town on Unguja island in particular. This might at a first glance seem rather nationally oriented but the geotags were not chosen out of a concern for Zanzibar as a place where a certain language variety is spoken that needs to be studied (even if that might ultimately be an interesting issue for research in World Englishes). They were rather employed for two reasons:

a) To create a manageable window of analysis on the digital tourist space

b) To create an unbiased window of analysis on the digital tourist space, as far as topics and language practices are concerned

The first issue refers to the potential size of the digital tourist space, which would be difficult to analyze in a meaningful way within a qualitative framework, at least within the scope of this paper. 30 posts were retrieved weekly from September 2019 until September 2020[2] *via* a dedicated API in Twitter, subsequently feeding R. This was necessary due to some difficulties with the Instagram API (see Lyons, 2019 for a more detailed discussion of this issue). Posts were retrieved from Stone Town and 5 kms around the city; an extension to other tourist locations where data for the study of the physical tourist space had been collected previously (Mohr, 2019; Mohr, 2020; Mohr, 2021) turned out to be futile due to a lack of social media activity there. Social media activity in general was limited, especially after the start of the COVID-19 pandemic and on a few occasions, there were fewer than 30 posts that could be retrieved per week.

The posts were subsequently "cleaned", i.e., three categories of posts were excluded from the data set. This refers to 1) posts that were not cross-posted on Instagram, 2) posts from accounts that are private and 3) posts that were not related to tourism in any way. With regard to 1), retrieving data *via* Twitter was necessary as mentioned above. However, posts that were only posted on Twitter and not on Instagram too, i.e., cross-posted, were not relevant for the study with its focus on Instagram. Concerning 3), the criterion was not interpreted very strictly so as not to restrict the data set too much. The determining factor were the participants themselves, i.e., participant profiles and previous posts were investigated closely in order to determine whether someone might have been a tourist in Zanzibar at the time of posting. Posts by hosts were in general easier to determine. The content of the post was another criterion considered. For instance, a post by a local Zanzibari, reading only "family" and showing a family picture, was excluded from the analysis. This overall procedure left a corpus of 363 posts (out of 534) in total. However, as indicated briefly above, the number and quality of posts changed considerably after the beginning of the COVID-19 pandemic, and only the data from 2019 are hence considered here. They comprise a corpus of 205 posts with 1,842 hashtags in total. The significance and use of hashtags are outlined in the next section of this article.

Issue b) above is related to an inductive approach to data collection, which is in line with ethnography in general. The application of geotags for data retrieval did not pre-determine the choice of participants, topics discussed by the participants of the tourist space, nor did it pre-determine language practices (including scripts or other semiotic material such as emojis) chosen in that space. Searching for particular key words would have been another option for data retrieval but since I wanted to treat the nature of topics forming part of tourism discourse and hence creating the tourist space as an empirical question, I

abandoned that idea. The use of key words would also have pre-determined the language practices I could have investigated and since the choice of these practices was another of the empirical questions to be answered by the study, I did not want to impact them. In this way, I applied an approach that was as inductive as possible and I pre-determined the tourist space as little as possible. This methodology then again emphasizes the close connection of place to the construction of space as suggested in transnational and translingual approaches.

Analytical Framework

Social media posts consist of many different components, drawing on diverse semiotic material (Zappavigna, 2015). In order to understand them in their entirety, all of these components have to be analyzed. Since this is the first study on digital tourism discourse associated with Zanzibar, choosing a certain focus was necessary for practical reasons. Hence, hashtags were chosen as an analytical framework based on their important functions on social media, linking different parts of the posts. Other material (text in the posts, pictures, likes) are briefly commented on where necessary. They will be the object of future analyses.

Hashtags have various functions that have been discussed extensively in the literature (e.g., Page, 2012; Zappavigna, 2015; Zappavigna and Martin, 2018). There are four functions that are relevant for this study. Others have put forth groupings into three main functions (e.g., Sykes, 2019) but a slightly more fine grained approach seemed more fruitful here. The first one is *supporting visibility and findability* (Page, 2012). Thus, posts can be found by other social media users interested in a specific topic and Instagram is indeed searchable for specific hashtags, which can also be followed. A user will employ a specific hashtag to be found by others interested in that topic, and possibly to generate likes. These have important phatic functions on social media in a similar way to hashtags. The visibility function is closely related to several other functions of hashtags that have been discussed in the literature. One of them is *folksonomic topic marking* emerging through community use (Zappavigna and Martin, 2018). This refers to users tagging their posts for what they perceive to be important issues in their posts. By doing so, they simultaneously *enact the ambient community* (Page, 2012), i.e., they create a subcommunity of, in this case Instagram, users interested in the same topic. Finally, they might express their *attitude towards that topic* or the post in general (in case it contains material they did not create themselves) and thus hashtags fulfill a meta-discursive function (Page 2012; Zappavigna, 2015). These four functions were central for the analysis provided in this article and for the coding scheme outlined in the next section.

Data Coding

The hashtags were coded according to several parameters using the NVivo software tool (QSR International Pty Ltd, 2020). The first parameter was language practices, i.e., English, French etc. The names for these practices were based on names that the participants in the investigation of the physical tourist space of

---

[2]There was a break in March 2020 due to the outbreak of the COVID-19 pandemic.

TABLE 1 | Language practices used in the hashtags.

Name of language practice	Number of hashtags
French	21
English-French	5
French-English	2
Russian	18
Spanish	13
Kiswahili	12
Turkish	12
Italian	9
Arabic	7
Portuguese	5
Malay	2
Japanese	1
Unclear	8
Total language practices other than English	**115**
English	1,727
Total all language practices	**1,842**

TABLE 2 | Semantic coding super-categories.

Coding category	Number of coding references
Geography and places	439
Travel	423
Identity	266
Nature	175
Feelings	130
Social media	108

Zanzibar used themselves (Mohr, 2019). They do not entail that I myself view them or other languages as bounded entities. For a discussion of the difficulties related to differences in epistemic interests and the choice of the "proper" ontology in sociolinguistics, see for example Røyneland (2020). **Table 1** shows an overview of the language practices used in the hashtags. As can be seen there, the majority of the 1,842 hashtags ($N = 1,727 = 93.8\%$) are in English. That means that the various and oftentimes large language repertoires of the Instagrammers are not utilized fully in hashtags in the digital tourist space of Zanzibar.

Secondly, the data was coded semantically. This resulted in 133 codes + 1 "other" category subsumed under 22 + 1 larger ("super") categories. The most frequent of these are shown in **Table 2**.

An example of the abovementioned codes and super-categories is "animals" with 69 coding references, which comprised the sub-categories "cats," "protection," "sea creatures" and "wildlife." Double coding was generally possible and necessary due to the complex nature of hashtags. An example of this is shown in (1), a hashtag which was coded for "feelings > positive", "geography and places > earth", "time" and "travel > vacation".

1) #beautifuldestinations_earthearthfocuswonderful_
   placesourplanetdaily

The third coding type referred to the hashtags' socio-pragmatic functions mentioned above, i.e., enacting the ambient community, expression of attitude, topic marking, supporting visibility/findability. Here, double coding was possible but rarer.

Like every method, this one had a few weaknesses. One issue that needs to be mentioned in this regard is that of naming language practices. While I do want to move away from conceptualizing languages as bounded entities, I did employ commonly used language names for coding. In this context, I

employed these names as coding categories only. Another problem that I encountered relates to the difficulty–if not impossibility–of clearly delineating languages and attributing linguistic material unequivocally to a particular language practice in superdiverse contexts. There are several words that are frequent in tourism discourse, and could be attributed to various language practices, among them *voyage*, which could be French or English for example, and *wanderlust*, which could be German or English. In these cases, contextual decisions were made, coding the material according to the language practice used mainly in the accompanying post. This is shown in example (2)[3], where *#wanderlust* was coded as English. Given the multilingual and translingual nature of the posts, these decisions are not completely incontrovertible though and eight cases that could not be determined remained (see **Table 1**).

2) Africa isn't all desert [...] #africa #tanzania [...] #travel **#wanderlust** #backpacker

The other problem that I experienced was my limited language knowledge. I retrieved linguistic material in languages that I do not speak (fluently), and determined the meaning of these posts and hashtags by using Google Translate. This generally worked well but I had to ask colleagues for their support in some cases. This specifically applied to material in what turned out to be Russian but was labelled as various Slavic languages by Google Translate. This emphasizes that working on projects like these makes interdisciplinary collaboration highly desirable.

RESULTS AND DISCUSSION

In the following, the results are presented according to the coding categories mentioned in the previous section, i.e., language practices chosen and other linguistic patterns observed in the hashtags (*Linguistic Patterns*) as well as themes observed, i.e., semantic categories (*Observed Themes*). A particular focus is on English in contact with other language practices, given the results shown in **Table 1**.

---

[3]A few hashtags have been deleted from this example in order to protect the Instagrammers privacy. This is indicated by [...].

Linguistic Patterns
General Patterns

Generally, there were two translingual linguistic patterns that could be observed. Either language practices were mixed within a single hashtag or they were switched across hashtags. Mixing within one hashtag was relatively infrequent though as shown in **Table 1**. Thus, only French and English were combined and only in 7 out of 1,842 hashtags. As can also be seen in the table, the pattern English + French was more frequent than French + English. The individual hashtags that could be found were (French material underlined): #travel*noire*, #travel*afrique*, #*Blackvoyageurs*, #*eurasietravel* and #*petitejoys*. Of these, #*Blackvoyageurs* made up almost half of all tokens. It was coded for "travel" and "identity" > "Black identity". Given that these two categories were among the most frequent super-categories, it is worth investigating these instances in more detail. #*Blackvoyageurs* occurred in four separate posts by two different Instagrammers. One of them had deactivated their profile at the time of writing this article. The anonymized post is shown in example (3).

3) Glowing different with [product name]
.
.
.
For a feature, follow and tag [account name] and use the hashtag #[account name]
.
.
.
You glow differently when you're on holiday✨
.
.
.
#womenwhoexplore   #weworktotravel   #passionpassport #blackvoyageurs #blackgirltravelslay
#iamtb   #btravelcreators   #travelnoire #blackgirltravelmovement #blackgirlstraveltoo
#blacktravelfeed #blacktravelmovement #blacktraveljourney #travelisthenewclub
#blackandabroad   #urbaneventsglobal   #mytravelcrush #weworktotravel #travelingwhileblack
#wanderlust #blacktravel #blackgirlstravel #melanintravel #blacktravelhackers
#blacktravelista   #workhardtravelwell   #soultravel #blacktravelgram #blackpassportstamps

In this example, Black identity seems to be one of the central issues to be emphasized by the hashtags #*blackvoyageurs*, #*blackgirltravelslay*,   #*btravelcreators*,   #*travelnoire*, #*blackgirltravelmovement*, #*blackgirlstraveltoo*, #*blacktravelfeed*, #*blacktravelmovement*, #*blacktraveljourney*, #*blackandabroad*, #*blacktravel*,   #*blackgirlstravel*,   #*melanintravel*, #*blacktravelhackers*,   #*blacktravelista*,   #*blacktravelgram*, #*blackpassportstamps*, as does female identity to some extent, as   expressed   in   the   hashtags   #*womenwhoexplore*, #*blackgirltravelslay*,   #*blackgirltravelmovement*,

#*blackgirlstraveltoo*, #*blackgirltravel*. All these hashtags together then create a sub-community on Instagram (the post had received 129 likes at the time of retrieval), i.e., Black female travelers, supported by one of the abovementioned functions of hashtags. Given its prominence for the authors of the posts themselves, which will be discussed further in the remainder of the analysis, this sub-group is reminiscent of a community of the imagination, which is formed as part of the larger affinity space. These communities are based on envisioned social relationships with others contributing to a strong sense of belonging (see also Wenger, 1998; King, 2019).

At the same time, some of these hashtags mark topics such as work and travel and support findability, especially a tag in -*gram* like #*blacktravelgram*, which is frequent (see *Ludic Tendencies and Linguistic Creativity* for further details). This function, as well as community creation, seems to be central to the post that has an advertising function. The use of *voyageurs* as a French word seems to support this, as the advertisement is for a beauty product. Further, the use of a language practice other than English appears to emphasize a certain air of worldliness that the Instagrammers want to convey and that seems important for the digital Zanzibari tourist space. This is also stressed by the use of the German loanword *wanderlust* in one of the hashtags.

The other posts using the hashtag #*Blackvoyageurs* emphasize all of the above arguments and stress pride in Black (female) identity, using for example hashtags like #*blackslayingit* [as shown in example (4)]. They were all posted by an organization "Blackvoyageurs" that has 153,000 followers at the time of writing and aims at advertising cultural and travel experiences around the world. Most posts show pictures of Black women, emphasizing their central role for that organization and the sub-community created. The issue of themes and various identities expressed *via* hashtags is discussed in more detail in *Observed Themes*.

4) Beach bum. So sad it's all come to an end. Could really be an island girl this place makes me so happy. Back to reality this time with a tan and with a light heart🧡
Corps de plage. Tellement triste que tout soit fini. Je pourrait vraiment être une fille des îles, cet endroit me rend si heureux. Retour à la réalité cette fois-ci avec un bronzage et un cœur léger😊
#Blackvoyageurs📍#Zanzibar #Tanzania
#melaninonfleek   #melanin   #afropolitan   #wanderlust #instatravel #blacknomad
#SurpriseTrip #AfricanAdventure #blacktravel
#blackking   #blackisbeautiful   #blackbeauty #melaninqueensonly #melaninart #blackrockingit
#africansthatslay #blackslayingit
#melanina #rainhanegra #negroelindo (📷[user name])

As can be seen at the end of example (4), where the camera emoji and a username are mentioned, the post was a re-posted picture from another user and hence the "Blackvoyageurs" were not in fact in Zanzibar when they posted. This illustrates the rather analytical nature of geotags, which might not be directly related to actual places in the world but rather create social spaces

TABLE 3 | Kiswahili hashtags used in the data.

Hashtag	Meaning
#hakunamatata	touristified Kiswahili: 'no problems,' general phatic marker
#jambo	touristified Kiswahili: 'hello,' general phatic marker or greeting
#lipapolepole	'pay slowly'
#sautizabusara	'voice of wisdom'
#tembeakenya	'walk Kenya'
#tukutaneporini	'meet in the wild'
#utaliikwawote	'tourism for all'
#UtaliiWaNdani	'local tourism'
#utaliiwetu	'our tourism'

FIGURE 1 | 30 most frequent hashtags ("#" removed).

TABLE 4 | *Insta-* and *-gram* hashtags in the data.

Hashtags beginning with *insta-*	Hashtags ending in *-gram*
#instacat	#blacktravelgram
#instafood	#catstagram
#instagay	#foodstagram
#instago	#gaystagram
#instagood	#MoodyGrams
#instagraffitiart	#petstagram
#instagrafite	#TagStaGram
#instalike	#travelgram
#instalove	#travelingram
#instamusic	
#instapassport	
#instapets	
#instasunsets	
#instatravel	
#instavacation	
#instawildlife	

TABLE 5 | Sub-categories of the "identity" category.

Coding category	Number of coding references
Average people	4
Black identity	51
Boys and men	8
Couples	3
Ethnic groups	1
Family	6
Girls and women	31
Globetrotter	13
LGBTQ+	24
Nationalities	8
Professional identity	60
Travel agencies	14
White identity	1

online. At the same time, this emphasizes the transnational nature of this social space, including not only participants from or in Zanzibar but also those who in some way feel affiliated with the archipelago and local tourism, i.e., this affinity space.

Another issue that is apparent in example (4) is the relatively frequent use of other language practices than English. This directly relates to the other linguistic pattern mentioned at the beginning of this section, namely that of switching language practices across hashtags. Thus, the hashtags above start in French, move to English with *#melaninonfleek* and to Portuguese with *#melanina*. Interestingly, both of these hashtags marking switches refer to skin color, a central concept for this group in terms of the identity that is indexed. Different language practices are also used in the posts themselves, English and French here. This emphasizes the diversity of the group writing but also the audience or public appealed to as do the multilingual hashtags, which can be found by Instagrammers using these other languages. The other two posts tagged *#Blackvoyageurs* were structured in the same way, with the English post preceding the French post and mixing different language practices in the hashtags. Transnationalism and

transnational identities are clearly invoked here, most importantly as social morphology and as hybrid cultural phenomena (Vertovec, 1999).

The Use of Kiswahili

Kiswahili as the official national language of Tanzania and Zanzibar deserves some attention. As can be seen in **Table 1**, there were only very few Kiswahili hashtags ($N = 12$). They are shown in more detail in **Table 3**.

The fact that there is little Kiswahili used in the hashtags is not surprising, given the general lack of fully fledged Kiswahili in the physical tourist space, especially among tourists. Tourists usually do not learn Kiswahili beyond a couple of formulaic phrases, and neither do migrants from outside of Tanzania and East Africa who come to work in the tourist industry (Mohr, 2020). Zanzibaris themselves use HMS, a pidginized form of Kiswahili, to communicate with tourists, which might be another reason why there is little fully fledged Kiswahili in the hashtags.

Of the hashtags mentioned in **Table 3**, *#hakunamatata* and *#jambo* occurred together, as did *#lipapolepole* and *#tembeakenya*, and *#utaliikwawote* and *#utaliiwetu*, thus

reducing the number of posts with Kiswahili hashtags to 9 (there are 4 tokens of #hakunamatata in total). All of these hashtags, except for #UtaliiWaNdani, which accompanied a Kiswahili post, accompanied English posts. This emphasizes the absence of Kiswahili and its token or exotic status in the digital tourist space of Zanzibar even more. Again, this pattern is not surprising given that only two of the posts were written by (local) tourism providers, while all others were written by tourists. The post tagged with #UtaliiWaNdani was written by a local tourist from mainland Tanzania, which explains the language choice. The presence of Kiswahili in the hashtags does certainly not contribute to the findability of the posts. Instead, it seems to stylize and exoticize the location, a function that has been attributed to indigenous languages in many tourist locations (see Salazar, 2006 on Tanzania). Their function is rather that of enacting and appealing to the ambient (tourist) community that is acquainted with a truncated but not a fully fledged version of Kiswahili. This emphasizes the importance of performativity, of putting on a certain identity in the digital tourist space, which is similarly central in the physical tourist space (Mohr, 2020).

Hakuna matata and jambo, part of HMS, create a rather touristified, pop-cultural, even "disneyfied" space online. They are central lexical items of HMS (Nassenstein, 2019) and hence very important for tourists. Two of the posts they accompany are especially noteworthy. They both include several pop cultural references as could be expected with these hashtags. These refer to the "Jambo song" and Walt Disney's "The Lion King" movie (Nassenstein, 2019; Mohr, 2020). The latter is explicitly mentioned in one of the posts [example (5)] and implicitly in the other [example (6)].

5) I'm sorry... I had to do it again!!!!
   People actually say it ALL the time...it's not just for #disney ●
   ●
   ●
   #hakunamatata #whatawonderfulphrase #handsome [...] #funny #sunshine #airport [...] #zanzibar #danzibar #tanzania #salsa #bachata #kizomba #singing #fun #friends

6) This Lion didn't sleep tonight!.
   .
   .
   .
   .
   .
   .
   #doingthesupermanthing #isolemlyswearthatiamuptonogood #livingmybestlife   [...]   #happy #thisisthedecadeivebeenwaitingfor   [...] #handstandsaroundafrica   #becauseiwasinverted   [...] #goodvibes #positiveenergy #hakunamatata [...] #travel #wanderlust   [...]   #handstands   [...] #handstandsaroundtheworld   #mischiefmanaged #upsidedown [...]

In example (5), there are several explicit references to Walt Disney, such as the hashtag #disney, and the song from the Lion King movie that made the expression hakuna matata so famous. Further, the pictorial material included in this post is a video showing the Instagrammer (a choreographer and life coach) and their group of friends, all starting to sing the Hakuna Matata song. This emphasizes that for most tourists, hakuna matata is associated with Walt Disney. Interestingly, they say in the post that "it is not just for #disney", thus wanting to emphasize the authenticity of the phrase and creating a type of insider identity with regard to Zanzibari and African culture (i.e., they enact the ambient community), possibly also stressing their worldliness. Absurdly, the post emphasizes exactly the opposite, as this pidginized form of Kiswahili is only used with tourists, and outsiders who do not know much about Zanzibari or Kiswahili culture. The lack of fully fledged Kiswahili in the presence of pidginized Kiswahili mostly used by tourists is reminiscent of Mock Spanish as described by Hill (1998). Mock Spanish refers to grammatically incorrect and/or hyperanglicized Spanish as used by White Americans in the South-West of the United States, often linked to particular stereotypes of Spanish speakers in that region. Similar phenomena have been reported by Blommaert and Backus (2011) with regard to various languages and their acquisition in the era of globalization. Some of the stereotypes described by Hill, e.g., laziness (emphasized by the frequent use of mañana in Mock Spanish), can also be found in HMS, e.g., the use of pole pole "slowly" as a general phatic marker. In contrast to the very explicit reference in (5), the pop cultural reference in example (6) is subtler and more ambiguous. "This lion didn't sleep tonight" could refer to the Disney movie but also to a popular pop song first performed in 1939 by Solomon Linda called "Mbube" ('lion') in isiZulu, before it was translated into English under the title "The lion sleeps tonight". The song was subsequently adapted and performed by several artists, one of the most common adaptations being the one by The Tokens from 1961. It is world famous and could very well be meant here. The other pop cultural references in the post are equally interesting. #isolemlyswearthatiamuptonogood and #mischiefmanaged refer to the Harry Potter Universe. These, together with #hakunamatata, and many of the other hashtags which are very positive as expressed by their coding "feelings > good," for instance #goodvibes and #positiveenergy, convey a positive attitude of the Instagrammer. This also emphasizes the attitudinal function of hashtags and the disneyfied, movie-like nature of the space created by this hashtag. Further, the post in example (6) is very playful, as is example (5), due to these hashtags and the accompanying picture, which included a handstand (not shown for privacy reasons). This is taken up in the hashtags #handstands, #handstandsaroundafrica and #handstandsaroundtheworld. Ludic tendencies like these are common in tourist contexts and are elaborated on in the following section.

Ludic Tendencies and Linguistic Creativity

Ludic tendencies, i.e., pop cultural references in the posts and in pictures and videos, are common in tourist contexts (Dann, 1996). They convey a laid-back tourist identity and in hashtags seem to enact the tourist community on social media, and hence have an important function. These hashtags

are also easily findable, by tourists and people interested in the pop cultural content referred to alike, and thus exhibit another central function of hashtags.

Ludic tendencies are closely linked to linguistic creativity and, as observed in hashtags, equally contribute to the enactment of the relaxed tourist community or affinity space. This creativity is visible in language play and specific word formation patterns. Here, the possibilities of multimodal meaning making become especially visible. Using upper and lower case letters in clever ways, tourists [see example (7)] and hosts [see example (8)] hence make use of the possibilities provided by script online.

7) #TANnedZANIA
8) #AGameOfTones

Both of these examples refer to tanning, a central activity among many tourists. This emphasizes the hashtag's function of enacting the ambient tourist community. This seems to extend beyond Zanzibar in example (8), as the hashtag was used by a tour provider that also offers services in mainland Tanzania, the Maldives, Australia and Europe; all of these destinations are also mentioned in the other hashtags accompanying the post in question. On the one hand, this illustrates that worldliness is a central concept for tourists that is catered for by hosts. On the other, it demonstrates that it might not only be a local, Zanzibari tourist community that is enacted with these hashtags but rather a global, transnational one.

In-group identity is also expressed by word formation processes such as compounding and blending. These are frequent with respect to semantic categories like Black identity, LGBTQ+ identity and traveler identity (three sub-groups and possible communities of the imagination), and often these formations include references to Instagram. Examples of the first were given earlier in *General Patterns* such as *#blacktravelgram*, examples of LGBTQ+ references are *#gaystagram, #instagay*, the latter also referring to Instagram and other examples being *#travelgram, #travelholic* and *#instavacation, #instatravel*. *Insta-* and *-gram* are extremely frequent in hashtags on Instagram and combine with all kinds of lexemes. **Table 4** shows all hashtags beginning with *insta-* or ending in *-gram* in the data.

These hashtags fulfil the function of enacting sub-communities or groups on Instagram centering around specific topics, the ones that emerge in **Table 4** seem to be the abovementioned Black and LGBTQ+ identities and travelers, as well as food, pet and specifically cat lovers. *#instagood* was especially frequent and is a bit different from the other hashtags in that it rather expresses an attitude towards a post. Linguistic creativity in marking in-groupness and the coining of new words is a commonly observed strategy in sociolinguistics and has also been documented for youth languages, for instance (Hollington and Nassenstein, 2015). The issue of identities performed and indexed by these patterns and the semantics of the hashtags are discussed in more detail in the following section.

Finally, another brief comment on the patterns observable in **Table 4** is necessary. Generally, formations in *insta-* were more frequent (both concerning types and tokens) than those in *-gram* but the reason for that remains unclear. From a structural perspective, these formations are blends. Given their high frequency, they seem to be an important in-group marker and an integral part of the language repertoires of specific sub-communities on Instagram.

Observed Themes

As shown in **Table 2**, there were six coding super-categories, i.e., larger themes, that were frequent in the data. Most of them are not surprising given the posts' concern with tourism: they referred to geographic locations, nature or traveling, as well as social media (e.g., the *insta-* and *-gram* formations mentioned above). This is also illustrated in **Figure 1**, showing the 30 most frequent hashtags in the data, where *#travelling, #zanzibar* and *#tanzania* are the top three. Other hashtags noteworthy here are *#sky* and *#sea*, which are hashtags belonging to the category of "nature" and *#instatravel* and *#instagram*, which were coded for the category "social media." The category "feeling," which was also part of the six most frequent super-categories, is represented by hashtags such as *#instagood* and *#love* in the word cloud.[4]

Identity was a frequently coded category; interestingly, it is only represented by *#globetrotter* in the word cloud. While this might imply that references to identity were not that frequent after all at a first glance, the lack of references among the 30 most frequent hashtags might rather suggest that the category as such is diverse. This is emphasized by **Table 5**, which shows an overview of its sub-categories and their frequency in the data.

As can be seen in **Table 5**, there are certain sub-categories of the "identity" category that are more frequent than others. These are "Black identity," "girls and women," "globetrotter," "LGBTQ+," "professional identity" and "travel agencies," each with more than 10 coding references. The frequency of categories related to traveling is again not surprising given that the study is concerned with tourism discourse. Some remarks are in order though, specifically concerning the globetrotter category, which included the word itself as well as *#worldtraveller*. These hashtags were exclusively used by tourists, some of them travel bloggers. Their posts included several other hashtags referring to traveling and (semi)professional identities, such as blogger. An example is shown in (9). In this case, the professional identity is also related to other functions of hashtags which are important for social media personae, such as *#follow* and *#like*. These hashtags are hence central for the performance of professional identities on Instagram.

9) The end of the road to Zanzibar 🌴 😍 TZ

#beautiful #paradise #zanzibar #tanzania [...] #travel #intrepid [...] #explore #africa #safari #visitafrica #wanderlust #adventure #backpacker #nomad #traveller

---

[4]While word clouds might not be a frequently employed method of visualization in academic papers, it was chosen here due to its quick readability and because it nicely captures the character of the data analyzed in this article, i.e., the playfulness of Instagram posts.

#livelife #worldtraveller #travelbug #blog #blogger #follow #like #share

Hashtags such as *#travel, #explore, #safari, #wanderlust, #adventure, #backpacker, #nomad, #traveller, #worldtraveller* and *#travelbug* in (9) all refer to the Instagrammer's identity as a traveler, as well as the appeal of traveling. This is emphasized by hashtags such as *#intrepid*, referring to desirable character traits for travelers, and *#livelife*, which stresses that traveling is a desirable activity that makes life worth living. Framing this with references to Zanzibar as a *#beautiful* location, even a *#paradise*, makes this identity even more desirable. The fact that no reference to nationality is made whatsoever (and nationality was in general not a frequent sub-category as shown in **Table 5**) seems to emphasize the Instagrammer's identity as an international traveler–on their travels, nationality seems to become less relevant. This is striking, given the fact that nationalities are very much relevant when it comes to crossing borders, as certain passports allow for easier travel than others. Not mentioning nationality hence might hint more at this and other Instagrammers not *needing* to worry about such matters and hence creates a certain kind of in-group as well. In any case, nationality is rarely expressed in hashtags and this emphasizes the transnational nature, via a reconstruction of place or locality, of the digital tourist space.[5] As mentioned above, other professional identities were also sometimes mentioned. Many of these are event planners, specifically wedding planners when referring to the poster's identity. Hashtags include *#destinationpartyplanner* and *#nigerianweddingplannerhouston*. While these hashtags certainly express identity and group membership, they also function as advertisements and thus serve visibility enhancing functions and perform the authors' professional role on Instagram. This is the case with many of the other professional identities mentioned, including the abovementioned travel agencies, as well. Professional identities of the people posted about, for example in the accompanying pictures, are an exception. They included for example *#fisherman* or *#musician*.

The other marked identities often referred to as shown in **Table 5** are Black identity, female identity and LGBTQ+ identities. These hashtags were additionally marked by using language practices other than English and especially creative hashtags, as outlined in the previous sections. These hashtags emphasize the special and sometimes minority status of these groups, at least with respect to the international travel community coming to Zanzibar. In contrast, the respective majority groups such as White travelers or men rarely indicate their (unmarked) identities by using specific hashtags (see **Table 5**). Further, all hashtags, except for *#boy*, in the category "boys and men" refer to LGBTQ+ identity as well. Examples are *#gayboy* and *#gayguy*. These seem to relate to, as mentioned in *General Patterns*, communities of the imagination, characterized by a strong

sense of belonging or in-group membership (Wenger, 1998), and thus a marked identity. This might be one of the reasons why other language practices than English are used in the hashtags creating these communities. For people from majority groups it seems to be less relevant to express their group membership, or they indicate membership in the larger tourist community (a majority group) and affinity space less closely connected than communities of the imagination (Gee, 2005; King, 2019).

What has been shown in this section is that besides indirect means of indicating identity, i.e. using particular language practices or linguistic creativity as outlined earlier in *Linguistic Patterns*, there is also the direct means of indicating one's identity by stating it in a hashtag. Neither of these means occur in isolation, they are usually combined in order for the poster to construct and perform their identity. One of the most characteristic examples of this is (4) above. The fact that all identities indicated in the examples provided here are multilayered, as illustrated for instance in (4) where Black and female identity are simultaneously invoked or in hashtags such as *#gayboy* where male and LGBTQ+ identities are indicated, supports previous accounts of complex identities online (e.g., Leppänen et al., 2020). It seems indeed as if roles can be mixed and matched (Turkle, 1995) in the online tourist space of Zanzibar (an affinity space), drawing on different social roles and groups (such as communities of the imagination) one feels affiliated with. The fact that many Instagrammers in the data invoked an identity as travelers also stresses that identities are temporary interactional positions (Bucholtz and Hall, 2005). The idea of traveling and the traveler inherently refer to temporariness and liminality (Mohr, 2020), as they can only be invoked in contrast to the notions of home and the people staying there. These identities are hence inherently temporary and fluid. Exceptions from this rule are the professional identities (of hosts) indicated by the hashtags, as well as the minority group affiliations mentioned above. Thus, the aforementioned Blackvoyageurs invoked their Black identity repeatedly, as did one Instagrammer from the LGBTQ+ community, using hashtags such as *#gay* and *#gayman* in all of their posts that form part of this data set. Membership in minority groups and expressing this clearly hence seems to be so central to their members that they constitute a relatively stable part of their performed identities online, even in a liminal space such as the tourist space.

CONCLUSION

The aim of this article was to provide insight into language choices and use in the digital tourist space of Zanzibar, with a special focus on identity construction through language. To this end, geotagged Instagram data was analyzed, using hashtags as a framework of analysis given their unique (meta)discursive functions. Hashtags and geotags provide a recent angle of analysis in the study of Englishes, moving away from theoretical and methodological nationalism (Schneider, 2019) and focusing more on the creation of space in line with a sociolinguistics of globalization (Blommaert, 2010).

---

[5]This was also mentioned by one of my participants for an earlier part of this study (see Mohr, 2020), who maintained that "people don't run around with flags on their heads," thus emphasizing that in the tourist space, nationality does not or should not matter.

Furthermore, correlating the data from the digital space with results from previous studies on the physical tourist space of Zanzibar provided insights into the nexus of offline and online communication, which is typical of contemporary communicative patterns (Blommaert, 2017). The results illustrate the polycentricity of communicative norms, amounting to invisible normativity (Hill, 1998; Blommaert, 2017), in a transnational social space. This is similar to what has been reported for the physical tourist space (Mohr, 2019; Mohr, 2020) and hence emphasizes the close connection between offline and online communication. The results also demonstrate the complexity of social spaces online, where various actors perform various identities, thus contributing to a loosely bound affinity space with different sub-groups or communities of the imagination with their own distinct linguistic practices (King, 2019).

The analysis reveals both similarities and differences concerning language choices and use in the digital as compared to the physical tourist space of Zanzibar. The absence of fully fledged Kiswahili is one of the most striking similarities. Thus, only a negligible percentage of the hashtags analyzed here were in Kiswahili, and the number of posts in Kiswahili was even lower. Much of the little Kiswahili that can be found is Hakuna Matata Swahili (Nassenstein, 2019), employed by tourists to signal their insider knowledge of Zanzibar and Zanzibari culture and supposedly to create an authentic space. This, however, actually demonstrates their lack of in depth linguistic and cultural knowledge. Kiswahili seems to fulfil an exoticizing function that is typical of indigenous languages in holiday locations and that has also been found in the physical Zanzibari tourist space. The use of HMS also expresses a laid-back attitude that many, specifically tourists in Zanzibar, are eager to invoke (Mohr, 2020). This is emphasized by the ludic linguistic tendencies in Kiswahili/HMS and especially English, that creatively make use of various semiotic means such as word formation patterns, typesetting and song (in videos), also observed in the data and typical of tourism discourse (Dann, 1996). The type of space that is constructed via the discourse is similar to the cartoonish imaginaries observed in the physical tourist space (Mohr, 2019; Mohr, 2020). These practices in particular are similar to Hill's (1998) findings on Mock Spanish and invisible normativity in that Kiswahili proper seems very much ignored, while cartoonish images of it are invoked and popular. In the Zanzibari tourist space, it seems to be the status as tourist and service-receiver, as well as host and service-provider that is the determining force concerning this particular language practice.

Transnational spatial patterns are created through language and specifically different language practices. In general, other language practices than English are rarer than in the physical tourist space where they are present in multilingual greetings specifically, making English less of a "multilingua franca" in hashtags online (Jenkins, 2015). This might be due to the functions of hashtags: to connect with the (international) tourist community and be visible, users employ hashtags that have been previously used or are likely to be found because they are in a language globally spoken. English is the hypercentral language of the world (de Swaan, 2002) and hence a good choice

in order to be visible and increase one's number of followers. Interestingly, choosing other language practices for hashtags (and posts) seems to express an identity relating to worldly tourists then, especially where French is concerned in some in-groups. Both the choice of other language practices and ludic tendencies were shown to be crucial for indicating membership in certain minority groups. Overall, deliberate language choices serve to index identity and (dis)affiliation in a diverse online space, emphasizing the need for investigations of speaker styles at the micro level in third wave sociolinguistics (see Eckert, 2012).

The combination of more indirect indexing of identities and social roles via language choices and particular linguistic structures as outlined above with direct indexing of identities through specific themes in hashtags shows that contemporary identities are multifaceted (Leppänen et al., 2020). The data further showed that identities in tourist spaces are indeed temporary interactional positions that can and will be abandoned after a relatively short time (Bucholtz and Hall, 2005). Minority group membership seems to be an exception, however. In this case, online identities seem to be an extension of offline identities, which further emphasizes the different character of these groups as compared to the much more loosely connected tourist space in general. Importantly, both more liminal and more stable identities in the context investigated here seem local and global at the same time (Leppänen et al., 2020) and this is one of the most interesting aspects of the nexus of offline and online identities in contemporary communication.

While the results presented here are interesting and could only be retrieved with the methods applied in this study, these had some weaknesses, as outlined in detail in *Methods and Data*. One issue that has not been addressed at length is the question as to who participates in the tourist space as it was analyzed here, i.e., who posts about Zanzibar on Instagram. As shown in the analysis, this includes people who were not in Zanzibar at the time of posting on Instagram. However, these were often representatives of large companies. One of my key participants from Zanzibar said that "Tour operators dont have skills of promoting through social media [...] They are few tour operators who use social media mainly facebook to attract DOMESTIC TOURISM rather than international tourists".[6] Thus, language choices and language use as observed in this data set seem skewed towards the tourist side of the tourist space, with hardly any local and smaller businesses participating. However, in the physical tourist space, they are a crucial part of the tourist sector (Mohr, 2021) and they should also be represented in studies of the digital tourist space. With the methods applied here, i.e., the API and subsequent observation, most hosts are relatively large and often international companies run by immigrants. In order to adequately account for smaller businesses, data from other social networks, especially Facebook, should hence be considered. Data collection to this end is currently being carried out.

The present article was the first based on the current data set and aimed at providing a preliminary insight into language practices in the digital Zanzibari tourist space. Hence, it only considered hashtags and only from 2019 for analysis. This leaves

---

[6]Quoted with permission from a WhatsApp conversation in June 2021.

several avenues for future research unexplored, among them an analysis of hashtags from 2020, which are also part of the larger data set, and a more comprehensive analysis of all components of the Instagram posts, including pictorial material. Especially the posts from 2020 and after the beginning of the COVID-19 pandemic will be an interesting basis for research into mobile language practices during a time of forced immobility and lack of travel. Analyzing speaker styles during these turbulent yet still times will certainly contribute to the study of language use and contact irrespective of national boundaries, and in the spirit of a sociolinguistics of globalization and transnationalism.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the author, within the constraints imposed by ethical considerations.

ETHICS STATEMENT

The data used in this article were anonymized to protect the privacy of the individuals concerned. Images could not be included for ethical reasons.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Al Zidjali, N. (2019). Society in Digital Contexts: New Modes of Identity and Community Construction. *Multilingua* 38 (4), 357–375.

Androutsopoulos, J. (2014). "Computer-mediated Communication and Linguistic Landscapes," in *Research Methods in Sociolinguistics: A Practical Guide*. Editors J. Holmes and K. Hazen (Oxford: Wiley), 74–90.

Androutsopoulos, J. (2008). Potentials and Limitations of Discourse-Centred Online Ethnography. Language@internet 5, article 8, https://www.languageatinternet.org/articles/2008/1610 (retrieved on August 9, 2019).

Barton, D., and Lee, C. (2013). *Language Online. Investigating Digital Texts and Practices*. London: Routledge.

Benor, S. B. (2010). Ethnolinguistic Repertoire: Shifting the Analytic Focus in Language and Ethnicity1. *J. Sociolinguistics* 14 (2), 159–183. doi:10.1111/j.1467-9841.2010.00440.x

Blommaert, J., and Backus, A. (2011). Repertoires Revisited: 'Knowing Language' in Superdiversity. *Working Pap. Urban Lang. Literacies* 67, 1–26.

Blommaert, J. (2017). *Durkheim and the Internet. On Sociolinguistics and Social Imagination*. London: Bloomsbury.

Blommaert, J. (2010). *The Sociolinguistics of Globalization*. Cambridge: Cambridge University Press.

Bolander, B. (2020). "World Englishes and Transnationalism," in *The Cambridge Handbook of World Englishes*. Editors D. Schreier, M. Hundt, and E. W. Schneider (Cambridge: Cambridge University Press), 676–701.

Bucholtz, M., and Hall, K. (2005). Identity and Interaction: A Sociocultural Linguistic Approach. *Discourse Stud.* 7 (4-5), 584–614. doi:10.1177/1461445605054407

Canagarajah, S. (2020). Transnational Work, Translingual Practices, and Interactional Sociolinguistics. *J. Sociolinguistics* 24 (5), 555–573. doi:10.1111/josl.12440

Chong, C. X. J. (2020). *Glossy Girls, Calm Girls and Grown-Up Girls: Comparing the Stylistic Practices of Young Singaporeans on Instagram* (Singapore: National University of Singapore). BA Honours Thesis.

Dann, G. M. S. (1996). *The Language of Tourism. A Sociolinguistic Perspective*. Oxon: CAB International.

de Swaan, A. (2002). *Words of the World: The Global Language System*. Cambridge et al: Polity.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2021). *Ethnologue: Languages of the World*. Twenty-Fourth Edition. Dallas, TX: SIL International. Available at: http://www.ethnologue.com (Accessed August 17, 2021).

Eckert, P. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annu. Rev. Anthropol.* 41, 87–100. doi:10.1146/annurev-anthro-092611-145828

Friedrich, P., and Diniz de Figueiredo, E. (2016). *The Sociolinguistics of Digital Englishes*. Oxon and New York: Routledge.

Gee, J. P. (2005). "Semiotic Social Spaces and Affinity Spaces: from the Age of Mythology to Today's Schools," in *Beyond Communities of Practice*. Editors D. Barton and K. Tusting (Cambridge: Cambridge University Press), 214–232. doi:10.1017/cbo9780511610554.012

Herring, S. C. (1996). *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*. Amsterdam: John Benjamins.

Heyd, T. (2016). "Global Varieties of English Gone Digital: Orthographic and Semantic Variation in Digital Nigerian Pidgin," in *English in Computer-Mediated Communication: Variation, Representation, and Meaning*. Editor L. Squires (Berlin: De Gruyter), 101–122. doi:10.1515/9783110490817-006

Hiippala, T., Hausmann, A., Tenkanen, H., and Toivonen, T. (2019). Exploring the Linguistic Landscape of Geotagged Social media Content in Urban Environments. *Digital Scholarship in the Humanities* 34 (2), 290–309. doi:10.1093/llc/fqy049

Hill, J. H. (1998). Language, Race, and White Public Space. *Am. Anthropologist* 100 (3), 680–689. doi:10.1525/aa.1998.100.3.680

Hollington, A., and Nassenstein, N. (2015). "Youth Language Practices in Africa as Creative Manifestations of Fluid Repertoires and Markers of Speakers' Social Identity," in *Youth Language Practices in Africa and beyond*. Editors N. Nassenstein and A. Hollington (Berlin and Boston: De Gruyter Mouton), 1–22.

Jenkins, J. (2015). Repositioning English and Multilingualism in English as a Lingua Franca. *English Pract.* 2 (3), 49–85. doi:10.1515/eip-2015-0003

King, B. W. (2019). *Communities of Practice in Language Research. A Critical Introduction*. London: Routledge.

Kubota, R. (2018). Unpacking Research and Practice in World Englishes and Second Language Acquisition. *World Englishes* 37 (1), 93–105. doi:10.1111/weng.12305

Leppänen, S., Kytölä, S., Westinen, E., and Peuronen, S. (2020). "Introduction. Social media Discourse, (Dis)identifications and Diversities," in *Social Media Discourse, (Dis)Identifications and Diversities*. Editors S. Leppänen, E. Westinen, and S. Kytölä (Oxon and New York: Routledge), 1–35.

Leppänen, S., Møller, J. S., Nørreby, T. R., Stæhr, A., and amd Kytölä, S. (2015). Authenticity, Normativity and Social media. *Discourse, Context and Media* 8, 1–5. doi:10.1016/j.dcm.2015.05.008

Lett, J. (1989). "Epilogue," in *Hosts and Guests: The Anthropology of Tourism*. Editor V. L. Smith (Philadelphia: University of Pennsylvania Press), 275–279.

Lim, L. (2020). "The Contribution of Language Contact to the Emergence of World Englishes," in *The Cambridge Handbook of World Englishes*. Editors D. Schreier, M. Hundt, and E. W. Schneider (Cambridge: Cambridge University Press), 72–98.

Lyons, K. (2019). Let's Get Phygital. *Ll* 5 (2), 179–197. doi:10.1075/ll.18025.lyo

Mair, Christian. (2020). "World Englishes in Cyberspace," in *The Cambridge Handbook of World Englishes*. Editors D. Schreier, M. Hundt, and E. W. Schneider (Cambridge: Cambridge University Press), 360–383.

Mair, C. (2018). "When All Englishes Are Everywhere: Media Globalization and its Implications for Corpus Linguistics and World English Studies," in *Anglistentag Regensburg (2017): Proceedings (XXXIX)*. Editors A.-J. Zwierlein, J. Petzold, K. Boehm, and M. Decker (Trier: Wissenschaftlicher Verlag Trier), 65–72.

Meierkord, C., and Schneider, E. W. (Editors) (2021). *World Englishes at the Grassroots* (Edinburgh: Edinburgh University Press).

Mohr, S. (2019). Assembling Concourse Material and Selecting Q-Samples on the Sociolinguistics of Tourism Discourse in Zanzibar. *Operant Subjectivity* 41 (1), 65–82. doi:10.15133/j.os.2019.005

Mohr, S. (2021). "English Language Learning Trajectories Among Zanzibaris Working in Tourism," in *World Englishes at the Grassroots*. Editors C. Meierkord and E. W. Schneider (Edinburgh: Edinburgh University Press), 70–90. doi:10.3366/edinburgh/9781474467551.003.0004

Mohr, S. (2020). Language Choices Among South African Migrants in the Tourist Space of Zanzibar. *South. Afr. Linguistics Appl. Lang. Stud.* 38 (1), 60–72. doi:10.2989/16073614.2020.1750966

Mohr, S. (2022). *Nominal Pluralization and Countability in African Varieties of English*. Oxon and New York: Routledge.

Nassenstein, N. (2019). "The Hakuna Matata Swahili: Linguistic Souvenirs from the Kenyan Coast," in *In Language and Tourism in Postcolonial Settings*. Editors A. Mietzner and A. Storch (Bristol: Channel View), 130–156. doi:10.21832/9781845416799-010

Onysko, A. (2021). "Where Are WEs Heading to? An Introduction to the Inaugural Volume of Bloomsbury Advances in World Englishes," in *Research Developments in World Englishes*. Editor A. Onysko (London: Bloomsbury), 1–10.

Page, R. (2012). The Linguistics of Self-Branding and Micro-celebrity in Twitter: The Role of Hashtags. *Discourse Commun.* 6 (2), 181–201. doi:10.1177/1750481312437441

QSR International Pty Ltd (2020). NVivo. Available at: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home (released in March, 2020).

Røyneland, U. (2020). "Regional Varieties in Norway Revisited," in *Intermediate Language Varieties: Koinai and Regional Standards in Europe*. Editors M. Cerruti and S. Tsiplakou (Amsterdam and Philadelphia: John Benjamins), 31–54.

Salazar, N. B. (2006). Touristifying Tanzania. *Ann. Tourism Res.* 33 (3), 833–852. doi:10.1016/j.annals.2006.03.017

Saraceni, M., and Jacob, C. (2021). "Decolonizing (World) Englishes," in *Research Developments in World Englishes*. Editor A. Onysko (London: Bloomsbury), 11–28. doi:10.5040/9781350167087.ch-002

Saraceni, M. (2015). *World Englishes: A Critical Analysis*. London: Bloomsbury.

Schmied, J. (2008). "East African English (Kenya, Uganda, Tanzania): Phonology," in *Varieties of English. Africa, South and Southeast Asia*. Editor R. Mesthrie (Berlin: Mouton de Guyter), 150–163.

Schneider, B. (2019). Methodological Nationalism in Linguistics. *Lang. Sci.* 76, 101169. doi:10.1016/j.langsci.2018.05.006

Shakir, M., and Deuber, D. (2019). A Multidimensional Analysis of Pakistani and U.S. English Blogs and Columns. *Eww* 40 (1), 1–23. doi:10.1075/eww.00020.sha

Sykes, J. M. (2019). Emergent Digital Discourses: What Can We Learn from Hashtags and Digital Games to Expand Learners' Second Language Repertoire. *Ann. Rev. Appl. Linguist* 39, 128–145. doi:10.1017/s0267190519000138

Tannen, D., and Trester, A. M. (Editors) (2013). *Discourse 2.0. Language and New Media* (Washington, DC: Georgetown University Press).

Thurlow, C., and Jaroski, V. (2020). ""Emoji Invasion": The Semiotic Ideologies of Language Endangerment in Multilingual News Discourse," in *Visualizing Digital Discourse*. Editors C. Thurlow, C. Dürscheid, and F. Diémoz (Berlin & Boston: De Gruyter Mouton), 45–64. doi:10.1515/9781501510113-003

Thurlow, C., and Jaworski, A. (2011). "Banal Globalization? Embodied Actions and Mediated Practices in Tourists' Online Photo Sharing," in *Digital Discourse. Language in the New Media*. Editors C. Thurlow and K. Mroczek (Oxford: Oxford University Press), 220–250. doi:10.1093/acprof:oso/9780199795437.003.0011

Thurlow, C., and Jaworski, A. (2017). Introducing Elite Discourse: The Rhetorics of Status, Privilege, and Power. *Social Semiotics* 27 (3), 243–254. doi:10.1080/10350330.2017.1301789

Tomlinson, J. (1999). *Globalization and Culture*. Cambridge: Polity Press.

Thurlow, C., and Mroczek, K. (Editors) (2011). *Digital Discourse. Language in the New Media* (Oxford: Oxford University Press).

Turkle, S. (1995). *Life on the Screen: Identity in the Age of Internet*. New York: Touchstone.

Vertovec, S. (1999). Conceiving and Researching Transnationalism. *Ethnic Racial Stud.* 22 (2), 447–462. doi:10.1080/014198799329558

Vertovec, S. (2007). Super-diversity and its Implications. *Ethnic Racial Stud.* 30 (6), 1024–1054. doi:10.1080/01419870701599465

Warner, M. (2002). *Publics and Counterpublics*. New York: Zone Books.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning and Identity*. Cambridge: Cambridge University Press.

Zappavigna, M., and Martin, J. R. (2018). #Communing Affiliation: Social Tagging as a Resource for Aligning Around Values in Social media. *Discourse, Context & Media* 22, 4–12. doi:10.1016/j.dcm.2017.08.001

Zappavigna, M. (2015). Searchable Talk: the Linguistic Functions of Hashtags. *Soc. Semiotics* 25 (3), 274–291. doi:10.1080/10350330.2014.996948

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

The Rising of "Alblish" (Albanian + English)—Data Collection and Analysis of Anglicisms in the Albanian Language

Irena Kapo*

Department of English Language, University of Tirana, Tirana, Albania

This paper investigates the impact of English on the Albanian language, a language contact phenomenon hitherto largely unexamined by Albanian and non-Albanian linguists alike. This contact mainly started as an indirect one at the beginning of the twentieth century, primarily via translations, the establishment of several educational institutions with the support of American and British associations and Albanian immigration into English-speaking countries. Several newspapers introduced a considerable number of Anglicisms into Albanian during the first half of the twentieth century, a good quantity of which "survived" even during the communist years of isolation. Those Anglicisms which made their way into the Albanian language until the 1990s and which have established themselves in the Albanian of today are mainly from the domains of culture and sports. The fall of the communist regime exposed the Albanian language to external linguistic and social factors, bringing about a flood of Anglicisms in different lexical fields other than culture and sport, such as business, information technology, politics, and medicine. The current database of Anglicisms in Albanian used for the purpose of this Brief Research Report was collected in the framework of the GLAD (Global Anglicism Database Network) project, http://gladnetwork.org. It includes 1,895 Anglicisms from newspapers and dictionaries, all integrated in a single file. The report first provides background information regarding the growing impact of English on the Albanian language during the three time periods under review (early twentieth century until 1960, 1960 to 1990, and 1990 to the present day). This is followed by an account of the data and the methods applied in the research. Subsequently, the findings and results are presented and discussed from a morphological, lexical, and syntactical perspective. The final outcomes are then used as a basis for the identification of Albanian language structures or patterns the English borrowings have managed to penetrate as well as of the level of adaptation English borrowings have in Albanian.

Keywords: Anglicisms, language contact, pseudo-Anglicisms, morpho-syntactic analysis, semantic fields, Anglicisms/English borrowings, language contact—English/Albanian

INTRODUCTION

The objectives of this brief research report are to provide evidence of how English has had an impact on Albanian from the beginning of the twentieth century, as well as to give an overview of the taxonomy of Anglicisms in the Albanian language.

For this purpose, the report focuses on English loanwords in the Albanian language. The main references used to determine the inclusion vs. exclusion of particular English borrowings/Anglicisms are the *Etymological Dictionary of the Albanian Language* by Kolec Topalli (2017), the four Albanian language dictionaries published by the Albanian Academy of Science (Cipo et al., 1954; Academy of Science of Albania, 1980, 1984, 2006), Manfred Görlach's *Dictionary of European Anglicisms* (Görlach, 2001) and Gaetano Rando's *Dizzionario degli Anglicismi nell' Italiano Postunitario* (Rando, 1987).

The indirect contact Albanian has had with the English language as well as British and American culture as a whole dates back to the beginning of the twentieth century, with numerous newspapers, translations from English as well as bilingual dictionaries published both within and outside the Albanian territory (D.J. Kamburi in 1915, K.A. Çekrezi in 1923, Nelo Dizdari in 1934, etc.—cited in Lloshi, 2011, p. 70). Among several other factors that contributed to this indirect language contact are the immigration of many Albanians into English-speaking countries such as Canada, the United States, England, or Australia and the establishment of several associations or cultural/sports clubs.

The impact of English on the Albanian language was considerably reduced during the approximately 50-year-period of communist isolation and intensified again only after the 1990s. The 1960s are taken as the first borderline as it was during this decade that the so-called "language purifying" reforms started to be established and implemented by the regime.

A few studies on Anglicisms in the Albanian language have been conducted in the twenty-first century so far, some of them as part of texts on the lexicology of the Albanian language, as newspaper articles on borrowings in the Albanian language (such as the writings of Shehu, 2002; Shkurtaj, 2004; Lloshi, 2005, 2011; Memisha, 2011) or as the subject of Ph.D. projects. The two most noteworthy and most wide-ranging of the Ph.D. theses conducted on the subject, in terms of the time periods covered as well as the corpus size under examination are that of Këçira (2005) and Kapo (2013). While the former extracted about 100 Anglicisms from a corpus of 4 years of the "Laboremus" newspaper (published in English and Albanian by the Albanian Vocational School, a technical high school established in Tirana by the American Red Cross in 1921), the latter extracted about 511 Anglicisms from 30 titles of different newspapers and magazines between 1990 and 2012. Both works conclude with observations regarding the phonetic, morphological and semantic aspects of English borrowings in the Albanian language.

Lloshi is one of the few Albanian linguists and perhaps the first who has underlined the fact that many of the recent borrowings used in the Albanian language have a much longer history in this language; he has also challenged statements of several other linguists that the Albanian language contains far fewer

Anglicisms than any other language or that the frequency of their use is limited (Lloshi, 2005, p. 32).

Albanian is one of the 16 European languages included in Manfred Görlach's *Dictionary of European Anglicisms* (DEA) which "is concentrated on the modern lexis imported after World War II, with a cut-off date of 1995" (DEA: xvi). Görlach underlines the absence of studies on Anglicisms in the Albanian language: "…Albanian had no literature worth mentioning" (Görlach, 2002, p. x). He only refers to two articles by linguists of the University of Pristina in Kosovo, I. Mehmeti and V. Nuhiu included in Rudolf Filipović's *The English Element in European Languages* (Filipović, 1982) as a basis for the Anglicisms used in the Albanian variety spoken in Kosovo. However, no studies or research projects on the impact of English on the Albanian language in Albania were taken into account. Therefore, the quantity and quality of the Anglicisms representing the Albanian language in the DEA is extremely limited. Notwithstanding the fact that Albania was not deeply explored as an area in this publication, Anglicisms have had an important place in Albanian for a long time.

METHOD AND CRITERIA USED FOR THE COLLECTION AND INCLUSION OF ANGLICISMS FOR THIS STUDY

About 90 newspaper titles from the three periods under review were consulted (the specific issues, with months and dates are cited in the Reference section) as well as the four Albanian Language Dictionaries published by the Academy of Science of Albania. These have been the source of the 1,895 Anglicism lemmas collected for the purpose of this study.

All the articles published in the newspapers were searched for Anglicisms with the titles as the main indicator for determining whether to read fully or simply screen the respective article, depending mainly on the topic and the source from which the information was taken. The fact that for the last 10 years these newspapers have been available online in digital format was of immense help during the research process. For research of earlier newspapers, they were physically retrieved from the fund of the Albanian National Library, consulted for the day and upon completion of work returned to the library staff within the same day. The English words and phrases identified in the newspaper articles were recorded alphabetically in Word files, together with the sentences they occurred in, the exact date, source, and page. If a word or phrase appeared more than three times, it was qualified as an eligible entry in the Anglicism lemma database built in an Excel file. The same procedure was followed for the Anglicism lemmas found in the four Albanian language dictionaries. The data was first alphabetically collected in a Word file and the entries were then incorporated in an Excel file with the purpose of collecting them for the GLAD Anglicism database [http://gladnetwork.org].

The Excel file of the GLAD database provides columns for the types of Anglicisms, the first date of attestation and the frequency of occurrence (carrying the *, **, or *** symbols to mark their

frequency). This organization of Anglicism lemmas facilitated the counting and the interpretation of the collected data.

The first criterion considered for the inclusion of an Anglicism found in our corpus is directly related to Onysko's definition of an Anglicism as: "any word recognized in its form (spelling, pronunciation, morphology) as coming from English" (Onysko, 2007, p. 90).

Secondly, the graphic criterion was decisive for the inclusion of the Anglicisms in the database. This criterion counts for those loans that have retained their English spelling and pronunciation in all or some of their morphemes/phonemes.

Henrik Gottlieb (2020, p. 41) describes the phenomenon of "relay borrowings" which is a scenario of language contact that is close to what happens to Anglicisms in Albanian:

"Language A exports an expression *a* to language B, thus coining expression *ab*. This is then borrowed in language C, resulting in expression *abc*."

After the World Wars "numerous translations and adaptations from the Romance languages like French, Italian as well as from German enriched the Greek-Latin portion of the modern Albanian language with a considerable number of neologisms" (Çabej, 1982, p. 122). Many schools were opened by Italian clergymen especially in the North of Albania; technology, commerce, finance and crafts were some other areas, apart from education, where Italian words entered the Albanian language in great numbers (Thomai, 1999, p. 226). This is also confirmed by Görlach (2002) who writes that "the dominance of Italian between 1916 and 1943 leads us to expect that English words would have been handed on through Italian" (p. 295). Këçira (2005), too, dedicates special attention to the English borrowings mediated in Albanian via the Italian language. Moreover, she introduces a list of Anglicisms mediated via the Italian language providing "a thematic ordering" in different semantic fields.

Thus, together with many Italianisms and French borrowings, many Anglicisms have been "transported" into Albanian via these two languages. Our corpus of Anglicisms shares 26% of the Anglicisms in the Italian language identified in Rando's *Dizzionario degli Anglicismi nell'Italiano Postunitario* (Rando, 1987).

For the purposes of this brief research report, the borrowings that have taken on distinctive features of the "transporting" languages (in our case, Italian or French) in their graphic form or in their pronunciation, have been excluded from our corpus of Anglicisms, for instance: [inovasion] for "innovation," or [golavaradʒ] for "goal average." Thus, the methodological concern related to the actual etymology of the loans collected is solved by retaining in the corpus those loanwords free from traces of French or Italian that have preserved English traces or indicators either in their graphic form or pronunciation, or extending to their semantic dimension.

At the same time, the corpus contains Greek- or Latin-based Anglicisms that can be listed as "internationalisms," as they have entered the Albanian language recently and mainly from English or American sources (such as the example of the loan "karizmë" (charisma), typical of the modern political discourse in Albania).

Another criterion used is the inclusion of semantic loans and loan translations in the database. This contributes to the final objective of this brief research report, i.e., identifying the degree of adaptation of English loans in the Albanian language.

RESULTS

This section presents some of the findings from the analysis conducted on the corpora of newspapers and dictionaries. First, the number and the percentage of the Anglicisms in the Dictionaries of the Albanian language will be shown. This includes the distribution of Anglicisms in the three time periods considered. A second subsection provides an analysis of the types of English borrowings in the Albanian language.

Diachronic Considerations and Frequency of Anglicisms in Albanian

In total, 31.18% of the Anglicisms of the database are collected from the Albanian language dictionaries. The number of Anglicisms in each of the dictionaries is given in **Table 1**.

Table 1 shows a rather constant volume of Anglicisms in Albanian from the 1950s to the 2000s. While the absolute numbers of Anglicism entries have been steadily increasing, the relative frequency measured on the total number of entries is actually the lowest in the most recent dictionary. To add to that, an increase of 47 Anglicisms from 1984 to 2006 is a rather small one.

The 1,895 Anglicism lemmas of the corpus include so-called "short-life Anglicisms," i.e. about 141 Anglicisms that, on average, have "lived and died" between 1910 and 1955 as well as so-called "revived Anglicisms," i.e., about 77 Anglicisms that entered the Albanian language during the first half of the twentieth century, "slept" during the isolation period, to be re-activated only after the 1990s.

The considerable number of Anglicisms, and all the constructions found, give an indication of how and how much the English language has permeated the Albanian journalistic discourse and entered the dictionaries. The following graph (**Figure 1**) illustrates the increasing trend of the use of Anglicisms in periods of 10 years, with the lowest numbers during the World Wars and during the isolation period.

If the time under investigation is divided into three periods, the graph in **Figure 2** emerges, with the two highest peaks in the 1920s, with 230 Anglicisms, and the second peak in the 2000s, with 258 Anglicisms. These numbers support the fluctuations in the Anglicisms intake of Albanian, before and after the isolation period.

Analysis of Types of English Borrowings in Albanian

The total number of Anglicisms extracted from the corpora is 1,895. **Figure 3** shows the distribution of the types of English borrowings in Albanian with 730 adapted borrowings (*xhins* for the English *jeans*, *diskurs* for the English *discourse*, etc.), 571 unadapted borrowings (*free lance, full time*, etc.), 132 loan translations (*jashtëshkollor* for the English *extracurricular,*

Anglicisms in the dictionaries of the Albanian language.

Dictionaries of Albanian and their year of publication	Number of Anglicisms/Total number of dictionary entries	Percentage of Anglicisms (%)
Dictionary of Albanian Language, 1954 (Cipo et al., 1954)	425/25,069	1.69
Dictionary of Modern Albanian Language, 1980 (Academy of Science of Albania, 1980)	540/41,000	1.31
Dictionary of Modern Albanian Language, 1984 (Academy of Science of Albania, 1984)	544/34,000	1.6
Dictionary of Albanian Language, 2006 (Academy of Science of Albania, 2006)	591/48,000	1.23

FIGURE 1 | Number of Anglicism lemmas per decade from 1900 to 2020 in the newspaper corpus.

FIGURE 2 | No. of Anglicism lemmas identified in the three periods.

FIGURE 3 | Types of borrowings.
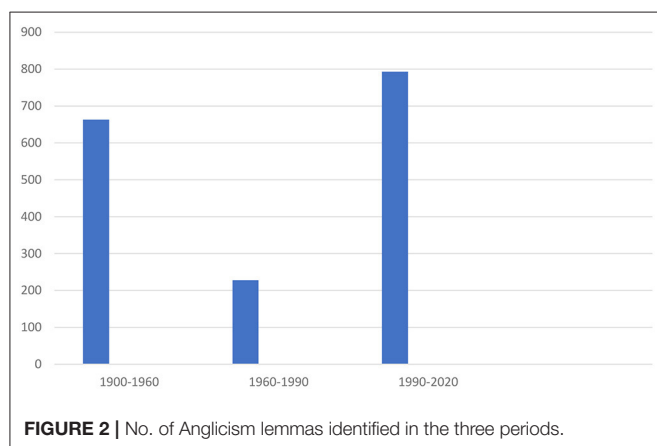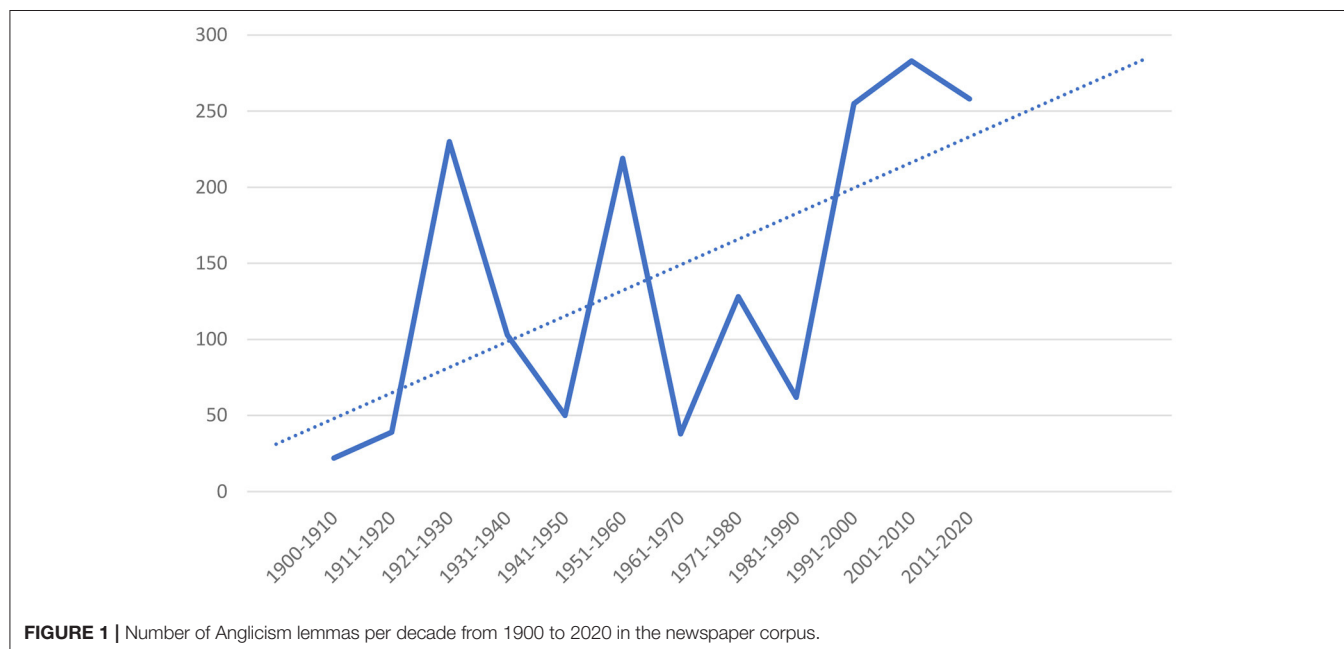
fundjavë for *weekened*, etc.), 480 hybrids (*printim*: *print* + the Albanian noun formation suffix *-im* for the English noun *print*, *gugëlloj*: *google* + the Albanian verb formation suffix *-oj* for the English *to google,* etc.), 105 pseudo-Anglicisms (*kondicioner* for the English *air conditioner*, *kritik* for the English *critical*, etc.), and 14 semantic loans (*profil, qasje, llogari,* etc. acquiring new meanings in Albanian under the impact of the English language).
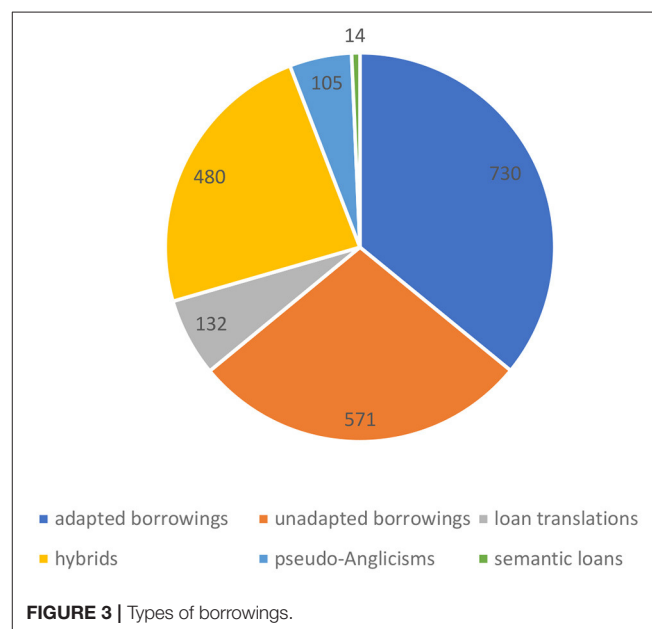
Figure 3 includes 141 "short-life Anglicisms" and 77 "revived Anglicisms."

The first group of "short-life Anglicisms" (7.4% of the corpus) includes entries such as:

1) the adapted borrowings *aktres* (actress) 1933–1943, *bek* (*back*—the player in the football game) 1930–1953, *egoistik* (egoistic) 1922–1942, etc.;

2) the unadapted borrowings *blond* (*blond*) 1928–1954, *round* (*round*—in wrestling) 1933–1949, etc.;

3) the hybrids *eksploatim* ("exploit+atim" for the English *exploration*) 1922–1952, etc., *eksprorijoj* ("expropriate+oj" for the English *to expropriate*) 1923–1933, etc.;

4) the pseudo-Anglicisms *centerfor* (the clipped *centerfor* for the English *center forward*) 1933–1964, *half* (for the English *half-back* in football) 1930–1953, etc.;

5) the loan translations *centërsulmues* (*center* – for the Albanian qendër +*sulmues*-forward from the English *centreforward*) 1938–1964, *topekambë* (top+e+kambë: *ball + and + foot* for the English *football*) 1943–1954, etc.

Examples in the second group, the "revived Anglicisms" (4.06%), recognize:

1) adapted Anglicisms such as the adjective *ekstrem* (for the English *extreme*) 1933–1945 and 1980-on, *bilion* (for the English *billion*) 1922–1938 and 1980-on, etc.;

2) the unadapted Anglicisms *fair play* (for the English *fair play*) 1922–1922 and 1996-on, *surplus* (for the English *surplus*) 1938–1938 and 2011 and on, etc.;

3) the pseudo-Anglicisms *flirt* (for the English noun *flirtation*) 1939–1939 and 1996-on, *ent* (for the English *entity*) 1926–1954 and 1990-on, etc.;

4) the hybrids *elaborim* (*elaborate+im* for the English *elaboration*) 1930–1933 and 1990-on, *atakoj* (*attack+oj* for the English verb *to attack*) 1927–1954 and 2013-on, etc.; and

5) the loan translations *liria e shtypit* (for the English *freedom of the press*) 1930–1955 and 1980-to date, etc.

As words and phrases of the Recipient Language coined with English constituents or inspired by English formations with no immediate counterpart in the source language, the pseudo-Anglicisms of the Albanian language constitute 5.5% of all the entries in the corpus. The following classification is conducted with reference to Gottlieb's typology and formations of pseudo-Anglicisms, based on Carstensen's German-based typology (Academy of Science of Albania, 1980), Furiassi (2010) regarding Italian and Fjeld's article on Norwegian lexicology (Fjeld, 2011):

1) "*clippings* (constituting the largest group of pseudo-Anglicisms in Italian) (Furiassi, 2010)

2) *recombination* of English morphemes

3) *neosemantization* of English lexemes (including semantic fossils and semantic neologisms" (Gottlieb, 2020, p. 208).

More than 70% of the pseudo-Anglicisms of our corpus can be categorized as belonging to the categories of "morphological change" or *clipping*. This category is common in many other European languages, as shown in Manfred Görlach's DEA (Görlach, 2001), affirmed also by Carstensen with the term "World-Wide Pseudo-English" (Carstensen, 1986, p. 831). Some examples still in use in current Albanian are: "exchange" (from *exchange office*), "glos" (from *lip gloss*), "happy end" (from *happy ending*), "ski" (from *skiing*), "nul" (from *null and* void), etc.

In the group of *recombinations* with English and Albanian constituents but with no counterpart in the source language, "bllokmen" is an example for a coinage typical for the Albanian language. "Block+men" refers to a "senior member of the government during the communist regime living in the area of *bllok*"—"the headquarters of the communist Albanian government in an isolated area near the center of Tirana." Other examples of *recombination* are "fotoreporter," "drogmen," "autostop," most probably via the impact of other languages.

The group of English pseudo-loans that has undergone a change in word class or that has developed a new sense in the Albanian language—the category of *neosemantization*—constitute 9.2% of the pseudo-Anglicisms collected. An example is "business lunch," used in Albanian as referring to "a special offer/discount restaurants make to people working in different institutions or organizations during their lunch break, usually fixed menus at a low price," and for the "change-of-word-class category:" "big up" used as a noun in Albanian.

There are also a few examples of spelling adaptations of English loan words in Albanian like: "ence" [entse] (for *hands*— the English *hands*—the offense of touching the ball with your hands in football, 1930–1950) and "xhem" [dʒem] (for *jump ball* in basketball, 1953–1989), two loanwords coined in Albanian from the "perceived pronunciation" of (mainly) non-English speaking Albanians regarding these two English sports terms. Nevertheless, the former is no longer in use while the latter is still used in colloquial speech.

From the 1,895 entries of the corpus, 1,305 are nouns, 252 adjectives, 173 verbs, 16 adverbs, and 4 are interjections. Most of the hybrids of the corpus, coined with English stems and Albanian word formation suffixes are fully integrated in the morphological system of the Albanian language, i.e., hybrid verbs are conjugated, and hybrid nouns take case, number, and gender inflections similar to the native words.

The number of compounds amounts to about 73 entries with different compound patterns, such as:

"noun + noun" compounds [*databazë* (for "database"), *fundjavë* (for "weekend"), *kameraman* (for "cameraman"), etc.]

"adjective + noun" compounds [*superyll* (for "superstar"), *toplistë* (for "toplist"), etc.].

The number of free phrases recognized in the database mainly as loan translations amounts to about 68, including noun and verb phrases such as "tavolinë e rrumbullakët" for *round-table*, "shoqëria civile" for *civil society*, etc.

In terms of word formation and the productivity of the English loans in Albanian, there appear to be about 253 word families or 16.6% of the corpus, from two to five constituents, coined with English constituents and Albanian suffixes: e.g., "lob–loboj–lobim–lobues" for the English "lobby (n.)–lobby (v.)–lobbyist": the English stem *lob* plus the Albanian verb formation suffix -*oj*, the Albanian noun formation suffix -*im,* and the Albanian adjective formation suffix -*ues*, etc.

FIGURE 4 | Semantic fields.
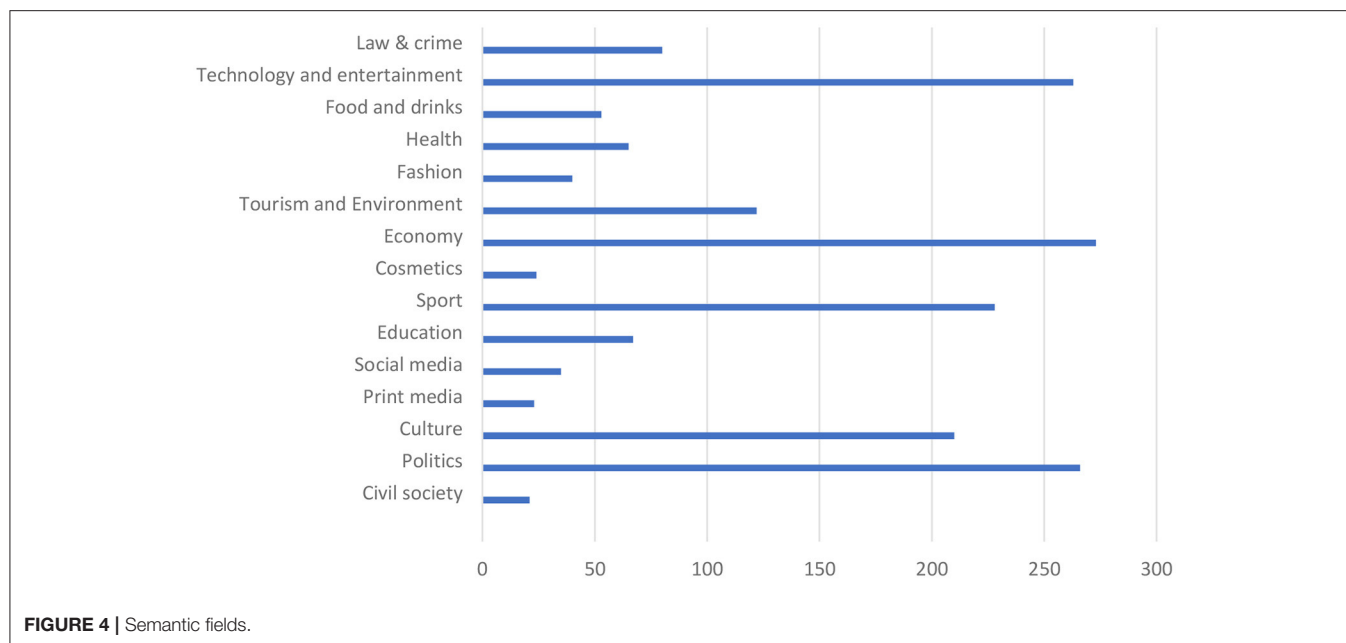
The English adjective *top* appears not only in unadapted compounds like *top-ten* or *top secret,* but it has become very productive with many other nouns in Albanian, such as "top-formacion" and "top-ekip" (for *top formation* and *top-team*— referring to football teams in particular), top-dokument (for *top-document*), top-blerje (for *the best buy*), etc.

The corpus includes 35 acronyms, most of which are unadapted in Albanian, both in their spelling and pronunciation (DJ pronounced ['di: dʒeɪ], CD [si: di:] DVD [di: vi: di:], GDP [dʒi: di'pi:]), etc.

The semantic fields to which the Anglicisms found in the corpus can be attributed are shown in **Figure 4**.

DISCUSSION OF RESULTS

This brief research report, built on a corpus of 1,895 Anglicisms in the Albanian language, sheds light on the non-steady intake of English loanwords in the Albanian language. The results show the influence of the communist regime on the use of Anglicisms in the written Albanian newspaper language with peaks before and after that. The results of the dictionary data, on the other hand, show a rather stable presence of Anglicisms.

Diachronic considerations on Anglicisms in Albanian are illustrated in **Figures 1**, **2**. The first period taken into consideration, 1900–1960, accounts for a number of Anglicisms that, in a way, formed the first layer of English borrowings in the Albanian press language. In total, 7.4% of the Anglicisms of the corpus "lived and died" during this first period, while 4.06% Anglicisms of the corpus "were born" during the first period, came out of use during the second period, and were "revived" in the third period.

The isolation years and the lack of external linguistic and extra-linguistic factors justify the lowest number of Anglicisms in the second period and the growing trend in the third period under consideration. After the second peak, that of the 2000s (with 258 Anglicisms)—as illustrated in **Figure 1**—we can speak of a "layer of English loans" adding to the Albanian lexicon, reinforcing the first layer established in the first period.

The data collected provide sufficient evidence for an analysis of the morphological and semantic features of English borrowings in the Albanian language.

As Haspelmath (2008, p. 7) states "Nouns are borrowed more easily than other parts of speech." In our corpus, the majority of Anglicism lemmas are nouns, with about 70% of all loan entries. The corpus of the Anglicisms collected from the newspapers shows that they can occur in various sentential positions and contexts. The majority of the nominal and adjectival English loans take inflectional and derivational endings of their respective gender, number, and case, according to the Albanian language rules.

An indicator of "good integration" is the occurrence of compounds with native elements (25.3 % of the corpus are hybrid formations). Referring to the numerous phrases, sentences, and contexts in which they appear, it can be affirmed that the majority of the English borrowings seems to have found their way into the Albanian language.

The percentage of unadapted borrowings is 30.7%. However, based on the number of examples retrieved for each of the Anglicisms of the corpus and on the rated frequency defined in the GLAD database, it can be estimated (even though rather subjectively) that from the unadapted Anglicisms of the corpus, 64% are very frequently and frequently used. Furthermore, about 35% of the adapted borrowings in the corpus, including loan translations, hybrids, pseudo-Anglicisms, and semantic

loans support the claim that there is a considerable degree of occurrence of Anglicisms in the Albanian newspapers consulted for the purpose of this research.

Onysko (2007, p. 52) regards the definition of pseudo-Anglicisms as "bordering on morphological and semantic changes of borrowings in the Recipient Language." As such, most of our pseudo-Anglicisms (about 70%) were coined via morphological changes and about 10% via semantic changes. The various semantic fields pseudo-Anglicisms belong to (sport, culture, technology, etc.), the English-influenced "new entries" in the Albanian lexicon as a result of "neosemantization" and "recombination" processes and their quantity of about 5% of the whole corpus is in line with Gottlieb's observation on pseudo-Anglicisms, namely that "they are rare but are useful" (Gottlieb, 2020, p. 204).

The criterion to include "loan translations" and "semantic loans" in the corpus served to test the integration of these two types of borrowings in the morphosyntactic structure of the Albanian language. From the cases investigated, about 4.2% of the corpus of our Anglicisms mark the initiation or the first stage of the loans' integration into the syntactic structures of the Albanian language. When a new meaning is introduced to the already existing lexemes in Albanian (the case of semantic loans) and when the English compounds or phrases are translated in Albanian (the case of loan translations), these two categories introduce neologisms and "new entries" to the Albanian lexicon: *burime njerëzore* (from "human resources"), *elefanti në dhomë* (from "the elephant in the room"), etc. Other neologisms introducing new notions, inventions, technological, cultural, culinary, etc. terms in Albanian are: "kompjuter" (for *computer*), "logo," "set," "sanduiç" (for *sandwich*), "link," "haker" (for *hacker*), "pas" (for *pass the ball* in football), etc.

Referring to the data provided in **Figure 4**, the semantic fields that top the list of the Anglicisms of our corpus are those from the areas of business, politics, information technology, sport, and culture.

CONCLUSION

The data and findings underscore the long-standing history of English and Albanian language contact and the oscillations this contact has been subject to since 1900 to date. The English loans collected for the purpose of this study were extracted from sources likely to reflect common language usage. Being only a manual micro-level analysis and with a few digitalized texts available, this research is limited regarding its provision of accurate numbers of English borrowings in terms of article hits and the share these Anglicisms have in relation to the total word number of the respective piece of writing. Thus, this study is by no means complete and conclusive.

In sum, the present study provides sufficient evidence and analysis for the starting point of an indirect English-Albanian language contact; it also provides some indicators regarding the linguistic "itinerary" English loan words have been taking into the Albanian language; the distribution of Anglicisms in various semantic fields, the "co-existence" with Albanian inflections, prefixes and suffixes or even as part of compounds and free phrases in Albanian point to the regular processes of how English loans become integrated in a range of RLs (without leading to deeper structural/grammatical changes of the RLs). Extensive code-switching during the last 10 years, in particular by the young generation but also by the media (subject to future research), further supports the rising of "Alblish" in the territory of Albania.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Academy of Science of Albania (1980). *Dictionary of Modern Albanian Language*. Tirana.

Academy of Science of Albania (1984). *Dictionary of Modern Albanian Language*. Tirana.

Academy of Science of Albania (2006). *Dictionary of Albanian Language*. Tirana.

Çabej, E. (1982). *Studime Etimologjike në Fushë të Shqipes, (Etymologycal Studies on the Albanian language), Vols. I, II, III, IV, V, VI, VII*. Tirana: Akademia e Shkencave e RPS të Shqipërisë (Academy of Science of the Socialist Republic of Albania).

Carstensen, B. (1986). "Euro English," in *Linguistics across Historical and Geographical Boundaries: In Honor of Jacek Fisiak. Vol. 2: Descriptive, Contrastive and Applied Linguistics*, eds D. Kastovsky and A. Szwedek (Berlin: Mouton de Gruyter), 827–835.

Cipo, K., Çabej, E., Domi, M., Krajni, A., and Myderrizi, O. (1954). *Dictionary of Albanian Language*. Tirana.

Filipović, R. (1982). *The English Element in European Languages*. Institute of Linguistics, Faculty of Philosophy, University of Zagreb.

Fjeld, R. V. (2011). *"Leksikalsk neografi i Norge,"* in *Nye ord. Sprognævnets Konferenceserie 1*, eds M. H. Andersen and J. N. Jensen (Copenhagen: Dansk Sprognævn), 19–39.

Furiassi, C. (2010). *False Anglicisms in Italian*. Milano: Polimetrica.

Görlach, M. (2001). *A Dictionary of European Anglicisms, A Usage Dictionary of Anglicisms in Sixteen European Languages*, Oxford: Oxford University Press.

Görlach, M. (2002). An Annotated Bibliography of European Anglicisms. Oxford: Oxford University Press.

Gottlieb, H. (2020). Echoes of English. Anglicisms in Minor Speech Communities – *with special Focus on Danish and Afrikaans*. Berlin: Peter Lang AG.

Haspelmath, M. (2008). "Loanword typology: steps toward a systematic cross-linguistic study of lexical borrowability," in *Aspects of Language Contact: New Theoretical, Methodological and Empirical Findings with Special Focus on Romancisation Processes*, eds T. Stolz, D. Bakker, and R. S. Palomo (Berlin; New York, NY: De Gruyter Mouton), 43–62.

Kapo, I. (2013). 'Monograph': Influence of English *Language* in the *Post-Communist Written Press* in Albania, *Dritan*.

Këçira, R. (2005). *Anglicisms in Albanian: A Direct and Mediated Language Contact Phenomenon*. Graz: Institut für Anglistik.

Lloshi, X. (2005). "Shqipja përballë rrjedhave të globalizmit (facing the globalization: the case of the Albanian language)," in Studia Albanica (Tirana: Academy of Sciences of Albania), 47–56.

Lloshi, X. (2011). *Raste Ndërhyrjesh të Huazimeve Nga Anglishtja (Cases of English Loans in Albanian language), Studime Filologjike – Phylological Studies. no. 1–2.*

Memisha, V. (2011). *Studime për Fjalën Shqipe (Studies on Albanian Language).* Qendra e Studimeve Albanologjike, Instituti i Gjuhësisë dhe i Letërsisë, (Center for Albanological Studies, Institute for Language and Literature), Tirana: Botart Publishing House.

Onysko, A. (2007). *Anglicisms in German. Borrowing, Lexical Productivity, and Written Codeswitching.* Berlin; New York, NY: Walter de Gruyter.

Rando, G. (1987). *'Dizzionario degli Anglicismi nell'Italiano Postunitario.* Firenze.

Shehu, H. (2002). *Rreth Anglicizmave në Mjetet e Sotme të Informimit Masiv, (On Anglicisms in the Modern Mass Media), Studime Filologjike, 3-4. Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë* (Academy of Science of Albania, Institute for Language and Literature).

Shkurtaj, G. (2004). Drejtshkrimi dhe Drejtshqiptimi në mediat e sotme të shkruara dhe të folura, (The correct writing and pronunciation in the modern written and spoken media). *Gjuha Jonë.*

Thomai, J. (1999). *Lexicology of the Albanian Language.* University of Tirana Publishing House.

Topalli, K. (2017). Etymological Dictionary of the Albanian *Language.* Academy of Albanological Studies.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

NEWSPAPERS AND MAGAZINES CONSULTED

1900–1960

Elcija i Zemres t'Jezu Krisctit, Kalendari, Korça, Drita, Hylli Dritës, Atdheu, Tomorri, Atdheu, Besa Shqyptare, Përparimi, Posta e Shqypnies, Kopështi letrar, Agimi, Adriatiku, Koha, Bagëti e Bujqësi, Fletore Zyrtare, Shqiptari Amerikës, Afrimi, Dajti, Republika, Laboremus, Demokratia, Revista Pedagogjike, Telegraf, Kombi Shqiptar, Gazeta Shqiptare, Gazeta e Re, Gazeta e Korçës, Arbëria, Besa, Gazeta e Tiranës, Vatra, Vullneti i Popullit, Cirka, Shtypi, Kombi, Bota e re, Balli Rinis, Bashkimi, Fletorja e Kryqit të Kuq Shqiptar, Sporti, Hosteni, Arsimi, Puna, Sporti popullor, Zëri i Rinisë, Zëri popullit, Ylli.

1960–1990

Ylli, Drita, Hosteni, Mësuesi, Sporti Popullor, Skena dhe Ekrani, Bashkimi

1990–2020

Sporti, Koha Jonë, Klan, Albania, Rilindja Demokratike, Shekulli, Gazeta Shqiptare, Republika, Jeta, Panorama, Korrieri, Spektër, Shqip, VIP, Studenti, Tema, Dita, Standard, Sport Ekspres, "55", Metropol, Monitor, Psikologji, Telegraf, GameOver, Mapo, Sot.

"We Use English But Not Like All the Time Like"—Discourse Marker *Like* in UAE English

Eliane Lorenz[1,2]*

[1]Department of English, English Linguistics, Justus-Liebig-University Giessen, Giessen, Germany, [2]Department of Teacher Education, English and Foreign Languages, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

The United Arab Emirates (UAE) is characterized by extensive language contact. Although Arabic is the official language, practically all communication in general as well as in higher education, in particular, takes place in English. The current study reports from the larger project *Language, Attitudes, and Repertoires in the Emirates* (LARES, 2019–2021) and investigates the use of English as a lingua franca (ELF) among university students in Sharjah, one of the seven sovereign emirates of the UAE. A spoken corpus based on 58 semi-structured interviews is used to examine the use of the discourse marker *like*. It has been shown to be a ubiquitous feature of English no longer confined to American English and occurs frequently in the corpus. It doubtlessly is a prominent discourse marker in the type of English spoken among the heterogeneous group of multicultural university students considered here. Although a large individual variation with respect to normalized frequencies of *like* can be observed, none of the social variables (i.e., gender, citizenship, L1, year of birth, number of languages, college, self-assessed proficiency in English, and English usage score) included in the analysis account for this variability. Instead, I argue that *like* as a discourse marker is part of the English repertoire of all students and appears to be even more frequently used than in other English varieties. This supports previous research arguing for an intensification of language change in ELF contexts as well as high individual variation as a characteristic of multilingual ELF users.

Keywords: discourse marker *like*, English as a lingua franca, spoken corpus, United Arab Emirates, varieties of English

INTRODUCTION

The current study is located in the United Arab Emirates (UAE) and investigates the use of the discourse marker *like* among university students in Sharjah. The focus is on the frequency of use as well as influence pertaining to social variables. First, in the two introductory sections, the status of English as a lingua franca (ELF) in the UAE is presented (*The United Arab Emirates and English as a Lingua Franca*), and a brief overview of the use of the discourse marker *like* in different varieties of English is given (*Discourse Marker Like*), which leads to the two research questions guiding this study. Second, in the *Methodology* section, the *Language, Attitudes, and Repertoires in the Emirates* (LARES, 2019–2021) project is introduced (*The LARES Project*), together with background information of the participants (*LARES Participants and Social Variables*) and the corpus as well as the data coding and analysis steps (*LARES Corpus: Data Coding and Analysis*). Third, in the *Results* section, the frequency of the discourse marker *like* is given (*Frequency of Like*) and contrasted

with the social variables of the participants (*Like Versus Social Variables*). In addition, the most frequent *like*-users are presented in some detail (*Most Frequent Like-Users*). Fourth, in the *Discussion* section, the current findings are discussed in light of previous research, in particular with respect to the overall frequency of *like* (*Frequency of Like*) and in conjunction with the social background information of the speakers (*Like Versus Social Variables*). Finally, the paper concludes with a short summary and outlook section.

The United Arab Emirates and English as a Lingua Franca

The UAE, located in the eastern part of the Arabian Peninsula bordering Saudi Arabia, Oman, and Qatar, constitutes a federation of seven sovereign emirates (Abu Dhabi, Ajman, Dubai, Fujairah, Ras al-Khaimah, Sharjah, and Umm al-Quwain). Even though the current study is set in Sharjah, it would be imprecise to neglect the other emirates because of their close geographical and political proximity. The seven emirates share the same constitution, they are strictly speaking not separated by borders, and the inhabitants have the same nationality (Siemund et al., 2020). For instance, it is not at all uncommon to live and work or study in two different emirates (Parra-Guinaldo and Lanteigne, 2021). Dubai is certainly the most famous of all sheikhdoms; however, Sharjah, along with the other emirates, shows a comparable development (Davidson, 2005; Madichie and Madichie, 2013; Siemund et al., 2020).

The metropolitan area of Sharjah is particularly fascinating for a (socio-)linguistic study because it is characterized by large-scale language contact and has undergone an interesting language development. Originally developing from small fishing villages (Pacione 2005; Siemund et al., 2020), the UAE has experienced unprecedented growth, mainly because of large-scale immigration. Ahmad (2016, p. 31) argued that the six Gulf Cooperation Council countries, to which the UAE belongs, "became an extremely attractive destination for skilled and unskilled labor from within the Arab World and beyond," particularly due to the oil industry. The UAE only gained independence from Great Britain in 1971, until which it had been part of a British protectorate, and has ever since been economically on the rise and gaining in popularity mainly because of the discovery of oil in the mid-twentieth century (Fussell, 2011). Another driving factor was the establishment and development of the tourism industry (Boyle, 2012; Leimgruber and Siemund, 2021). Whereas in 1971 population figures were below 300,000, they have, 50 years later in 2021, reached over 10 million, which is more than a 33-fold increase.[1]

Moreover, what is particularly intriguing about the UAE is that the nonnational population greatly outnumbers the local Emiratis in all seven emirates. For example, there are only about 10%–15% Emirati nationals living in Dubai (Government of Dubai, 2019; Dubai Population, 2020) and approximately 12%

Emirati inhabitants in Sharjah (Sharjah Population, 2020). The nonnational population represents a diverse multilingual and multicultural group, with South Asians being the largest (approximately 60% of the entire population in the UAE) (Al-Issa, 2021). Thus, the UAE is what Vertovec (2007) called a society characterized by super-diversity (see also Hopkyns, 2021). This high share of expatriates in the UAE has essentially helped to guarantee its fast economic growth, as most of those coming to the UAE are not refugees but economic migrants (Al-Issa, 2021).

This population distribution ultimately creates a complex linguistic landscape (Boyle, 2011) and necessarily results in intense language contact at all levels of society (Siemund et al., 2020; Siemund and Leimgruber, 2021; Siemund, 2022). Al-Issa (2021) reported that there are more than 100 languages represented in the UAE, which are, apart from Arabic and English (which are listed as "principal languages" on *Ethnologue*[2]), for example, Bengali, French, Farsi, Hindi, Malayalam, Pashto, Punjabi, Somali, Tagalog, Telugu, and Urdu. Arabic is the official language of the UAE, but in order to work or live in this country, it is strictly speaking not necessary to have a command of Arabic (Al-Issa, 2021). Instead, it is competence in English, which ensures successful communication and secures job opportunities. For instance, anecdotal evidence presented in Hopkyns (2017, 2021) showed that not being sufficiently proficient in English "could be seen as a linguistic disability," even for daily tasks such as going shopping (Hopkyns, 2021, p. 253). Apart from the Arab expatriates, other migrants rarely use Arabic but heavily rely on English as the language of communication. English has also become more important among Emirati citizens, especially among the younger generations, as many families employ nannies with whom they communicate in English (Hopkyns, 2021). The important and ubiquitous role of English can also be supported with a quote from one of the participants of the current study, who is an Emirati citizen:

> (1) *Because as I said we live in a very diverse place I think English is one of the connecting languages that we have* (f8)[3]

Not only are there many different languages present, but there are also different native and non-native varieties of English spoken in the UAE (Parra-Guinaldo and Lanteigne, 2021; Thomas, 2021). Many of the numerous expatriates grew up in countries where English is at least one of the official or national languages (such as India or the Philippines) or come from countries where English is the majority language, such as the United Kingdom or the United States (Hopkyns, 2017). Therefore, many citizens have been in contact with English

---

[1]These numbers were taken from https://www.worldometers.info/world-population/united-arab-emirates-population/, accessed July 28, 2021.

[2]See https://www.ethnologue.com/country/AE.

[3]This quote comes from the LARES corpus, which will be presented in *LARES Corpus: Data Coding and Analysis*. The speakers remain anonymous and were numbered randomly. The only information visible from the ID is their gender. The letter "f" stands for female and "m" for male. All further quotes are also taken from the LARES corpus.

during school education in a foreign country, often following an English curriculum. Others have studied English as a foreign language (EFL) in school. In the UAE, English is introduced as an obligatory school subject already from the early school years onwards; in addition, there are many private schools that have English as their medium of instruction, and practically all higher education takes place in English (Hopkyns, 2017; Al-Issa, 2021; Thomas, 2021). Some Emirati families even decide to send their children to private instead of government schools where the medium of instruction is English (Ahmed, 2021). This is motivated because in order to be admitted to a university, an entry test documenting sufficient English proficiency is usually a requirement (Ahmed, 2021; Al-Issa, 2021). Interestingly, similar placement tests for Arabic do not exist (Ahmed, 2021). It clearly follows that whoever wants to pursue a university degree in the UAE needs to know English (but not necessarily Arabic).

Typically, English is associated with modernity and internationalism, used for business and education, and with this, it "dominates everyday public life and, to a lesser extent, private life too" (Hopkyns, 2017, p. 40; see also Thomas, 2021, for a recent overview of English users and the use of English in the UAE). Arabic, however, is largely confined to the home and family context as well as to practicing religion (Al-Issa, 2021). Without any doubt, English has developed into the lingua franca (Theodoropoulou, 2021; Thomas, 2021), and it has replaced Arabic in many domains (Fussell, 2011; Al-Issa, 2021). ELF could be defined as a vehicular language used by speakers who do not share the same language (Filppula et al., 2017; Mauranen, 2012, 2017). As has been explained above, the result is a complex contact situation of diverse multilingual speakers. It is precisely in such situations of multilingual contact that new varieties of English will emerge (Mair, 2021).

Some have tentatively argued that a new variety of English has emerged or will emerge in the UAE, referred to as either "Gulf English" (Fussell, 2011, p. 31) or "UAE English" (Boyle, 2012, p. 321). Gulf English, clearly not confined to the UAE but to be found in the larger Gulf region, is said to have emerged because of language contact between speakers of Arabic and expatriates who speak different varieties of English (Fussel, 2011). According to Fussel (2011), in the early 2000s, this variety was still at an initial stage, moving towards Schneider's (2007) phase three of the Dynamic Model, i.e., the nativization phase. Boyle hypothesized shortly after Fussel (2011) that the English found in the UAE "should in time become 'UAE English', a variety, perhaps, with a distinct South Asian flavour" (2012, p. 321). Others, however, remark approximately 10 years later that a "local norm of Dubai or Gulf English" has not yet emerged and that "it remains to be seen if it will ever develop" (Leimgruber and Siemund, 2021, p. 1; see also; Ahmed, 2021; Siemund et al., 2021).

As there is to date a lack of research investigating the use of this lingua franca and its status as a new English variety (Siemund et al., 2020), the proposed study sets out to examine the use of ELF in the UAE. There are a number of recent linguistic studies based on Dubai, Sharjah, or the UAE in general (Randall and Samimi, 2010; Boyle, 2011; O'Neill, 2014; Thomas, 2016; Cook, 2017;

Piller, 2017; Parra-Guinaldo and Lanteigne, 2021). However, most of these are based on a limited number of participants or specific subgroups (such as female students, see O'Neill, 2014; or police officers, see Randall and Samimi, 2010) or on a small set of linguistics examples that do not yet qualify to formulate generalizations (Boyle, 2011; Parra-Guinaldo and Lanteigne, 2021). The current study is of course also based on a limited number of participants, representing only one group of the entire population (see *Like Versus Social Variables*); yet this research tries to add a puzzle piece to the emerging picture and to contribute to studies investigating varieties of English.

More specifically, this study examines the use of the discourse marker *like*. In general, *like* has received much scholarly attention (see *Discourse Marker Like*), and it is clearly undergoing frequency shifts in present-day Englishes (D'Arcy, 2017). It has mainly been studied in native Englishes, but much less research focuses on non-native speakers of English or ELF varieties (see, for example, Diskin-Holdaway, 2021 or; Rüdiger, 2021). Yet studies investigating language change in ELF varieties identified accelerated grammatical language change (Laitinen, 2020). The observed frequency shifts in present-day Englishes may be understood as language change of a pragmatic phenomenon. Therefore, the current research aims to add to this latter context by investigating the discourse marker *like* used in the English spoken in the UAE. The following subsection introduces the discourse marker *like* and its use in different varieties of English.

Discourse Marker *Like*

This section focuses on an extremely versatile and multifunctional word form in English. The four-letter word *like* has been shown to appear with (at least) 12 different functions. D'Arcy's (2017) account of this word form, which is arguably the most comprehensive, distinguishes between the "unremarkable" functions as a verb (*I like ice cream*), adjective (*they are as like as twin brothers*), noun (*such as fishing or the like*), preposition (*a difficulty like this*), conjunction (*like I said*), complementizer (*it feels like a bit too much*), and suffix (*an Earth-like planet*). D'Arcy (2017) further listed some more notable functions of *like* as an approximative adverb (*it took like 3 hours*), a sentence adverb,[4] a quotative use (*and I was like . . .*), a discourse marker (*like they accepted it*), and as a discourse particle (*he was like falling asleep*).

For the current study, only the last two uses of *like* are of interest and will be focused on in the following sections. D'Arcy (2017, p. 14) stated that "as a discourse marker, *like* encodes textual relations by relating the current utterance to prior discourse" and that it "signals exemplification, illustration, elaboration, or clarification." She further points out that its use increased, particularly so in the second half of the last century, and that it is "widely attested across varieties of English in speech

---

[4] *Like* as a sentence adverb seems to be restricted to some dialects of English, particularly to English dialects in Ireland (D'Arcy 2017, pp. 12–13). D'Arcy (2017, p. 13) provides some examples of this use, for instance, "You'd hit the mud on the bottom *like*."

materials" (D'Arcy, 2017, p. 14). In opposition, D'Arcy (2017, p. 15) explained that particularly as a particle, *like* "signals subjective information" and establishes "common ground, solidarity, or intimacy" between the speech partners. It is this latter use of *like* that is prominently associated with young and female speakers, driving the observed frequency shift mentioned above. Examples (2) and (3) exemplify these two uses respectively. Notice that as a discourse marker, *like* occupies the clause initial position, yet as a particle, it occurs clause internally.

(2) *Mostly the US. Like all the states in the US.* (f23) [discourse marker]

(3) *It wasn't like a challenge but still it was something new for me.* (m4) [discourse particle]

Other than D'Arcy (2017), the current study follows Schweinberger, who does not distinguish between discourse marker and discourse particle uses but who refers to both types using the label "discourse marker" (2014, p. 52). The remaining discussion also refers to both types (which are further subclassified according to their position, see *LARES Corpus: Data Coding and Analysis*) as discourse marker *like*. This is justified because both marker and particle uses (as defined by D'Arcy 2017) of *like* share a number of features outlined below.

An important property of discourse markers is their optionality. This means that they are not required for a sentence to be grammatical (Fuller, 2003). Moreover, particularly in interviews, the use of discourse markers fulfills a stylistic role by creating a rather casual style. This seems to be reinforced by their co-occurrence with other markers; for example, *well* and *let's see* and could be understood as an "interactional tool" to establish a common basis between the interlocutors (Fuller, 2003, p. 372). Fuller (2003, p. 370) further argued that *like* is "pragmatically useful" for creating closeness, placing focus on something, or implying approximation, which are typical contexts of (personal and somewhat informal) interviews. Even though *like* (among other discourse markers) is a way to accommodate planning and continuation in spontaneous speech (Hasselgren 2002; Wolk et al., 2021), it is frequently considered as something negative. Rüdiger (2021, p. 1) even talked about "public language stigmatization" and presents a number of quite strong, negative attitudes towards the use of *like* (Rüdiger, 2021, p. 2).

A number of recent publications focus on the use of the discourse marker *like* (e.g., Schweinberger, 2014; D'Arcy, 2017; Diskin, 2017; Gabrys, 2017; Corrigan and Diskin, 2019; Corrigan and Diskin, 2019), most likely because it is such a prominent or salient feature of English across its different varieties and due to its frequent use (Schweinberger, 2014; Corrigan and Diskin, 2019; Leuckert and Rüdiger, 2021). Yet it can also undergo frequency shifts (D'Arcy, 2017), and its frequency of use differs across English varieties, as impressively demonstrated in Schweinberger's (2014) comprehensive study. To provide only a selection, Schweinberger (2014, p. 185, 379) reported frequencies of the discourse marker *like* ranging from 0.49 per

one thousand words (ptw) for British English, 1.51 ptw for Indian English, 2.18 ptw for New Zealand English, and 2.23 ptw for Philippine English, up to 4.38 ptw for Canadian English. These differences underline the importance of investigating the use of *like* in other English varieties.

Whereas many studies focus on native speakers of English, fewer studies target non-native speakers of English (such as L2 or foreign language users) or ELF varieties (however, there seems to be an increasing interest in analyzing discourse markers among second or foreign language learners, see, for example, Gilquin, 2016). Keeping in mind the frequency differences across English varieties discovered in Schweinberger (2014) and the fact that, typically, the use of discourse markers such as *like* are not normally taught to foreign language learners in schools (Mukherjee and Rohrbach, 2006, p. 216; Rüdiger, 2021, p. 2; Wolk et al., 2021, p. 10), the use of *like* among non-native speakers proves particularly insightful. Moreover, non-native speakers of English are often found to use discourse markers less frequently and differently than native speakers (Liao, 2009; Gilquin, 2016). Mukherjee and Rohrbach (2006) provided an explanation for this. They argue that discourse markers are among the more challenging elements when learning a foreign language, and thus, they are typically acquired relatively late (Mukherjee and Rohrbach, 2006, p. 213). Yet the (correct) use of discourse markers contributes to sound more native-like or, in other words, more natural or idiomatic (Wolk et al., 2021, p. 9; see also Liao (2009)). The following selected findings of studies about discourse markers in general as well as *like* in particular underline that more research focusing on non-native speakers of English is needed.

Hasselgren (2002), for instance, discovered that higher proficiency, or rather higher fluency in English, resulted in a more target-like use of what she called "smallwords." The discourse marker *like* was among the smallwords investigated in her study comparing native speakers of English with Norwegian learners of English, grouped into more fluent and less fluent users of English (Hasselgren, 2002). *Like* appeared among those smallwords, which were acquired comparably late and thus require a certain level of proficiency. Similarly, in a study on pragmatic markers, Neary-Sundquist (2014) found that with increasing English proficiency of Korean and Chinese L2 learners, the use of these pragmatic markers increased as well. This conclusion was based on overall frequencies, and no specific mention of the discourse marker *like* was made, even though *like* had been part of the analysis. Gilquin (2016) also underlined the importance of proficiency, but in particular language skills acquired through naturalistic language exposure. More precisely, Gilquin (2016, p. 216) noted that for non-native learners to acquire the use of discourse markers, "exposure to naturalistic speech outside the classroom" is particularly important. Moreover, she found that native speakers of English (United Kingdom) used the discourse marker *like* approximately 3.5 times more frequently than non-native speakers of English (various L1s) (Gilquin, 2016, p. 220). In addition, she reported a statistically significant difference between those foreign language learners who had spent some time in an English-speaking country versus those who had not, with the

former using *like* more frequently than the latter (Gilquin, 2016, p. 221). She further exemplified that with the increasing length of stay (particularly so after 10 months), the frequency of *like* increased (Gilquin 2016, p. 227). This underlines the importance of exposure to language use outside of normative EFL contexts, where the (over)use of discourse markers would rather be discouraged, in order to acquire this discourse marker. Another striking finding was the observation that some foreign language learner groups used *like* relatively frequently, whereas others barely produced it at all. Among the former were the Polish, Dutch, and Swedish, as well as the German and Spanish learners of English, who had at least some (naturalistic) exposure to English in their respective countries of origin, for instance, *via* media or the internet. Chinese, French, and Italian speakers of English, who were assumed to have more limited access to naturalistic English language, were among the latter group. Gilquin (2016) further predicted that English as a second language (ESL) speakers used discourse markers more frequently than EFL speakers. This could be confirmed with *like* occurring more frequently in the ESL than in the EFL data (Gilquin, 2016, pp. 240–241).

Müller (2005) also identified frequency differences between native and non-native speakers of English, namely, that the native speakers of English (US) used the discourse marker *like* more frequently than the foreign language learners of English (L1 German). More precisely, all American speakers used the discourse marker *like* at least once, whereas only less than 60% of the German learners of English had at least one occurrence of *like* in their utterances (Müller, 2005, p. 230). Furthermore, she found some age-related tendencies, namely, that the younger German learners of English used the discourse marker *like* more frequently than those in the middle-age group (Müller, 2005, p. 232). The speaker relationship is also shown to have an effect. Among friends, *like* was used more often than among strangers (Müller, 2005, p. 233). In addition, Müller showed that interaction in English in informal situations as well as using English as the primary means of communication at least occasionally had boosting effects on the use of discourse marker *like* (Müller, 2005, p. 239). Finally, the influence of American English was prominent in her study. She showed that those German learners of English who had spent time abroad in the United States had higher rates of *like* than those who had been to the United Kingdom (Müller, 2005, p. 239).

In a study by Wolk et al. (2021), *like* was the fifth most frequently used discourse marker, but it generally appeared relatively infrequently (Wolk et al., 2021, p. 23). Their findings are based on English major university students with German, Spanish, Bulgarian, or Japanese as native languages. They further noticed that the L1 background influenced the use of discourse markers. For example, *like* appeared most frequently among the Spanish students studying in Madrid (Wolk et al., 2021). Moreover, they could also show that length of English instruction positively correlated with the frequency of discourse marker usage (Wolk et al., 2021, p. 30).

Interestingly, Diskin-Holdaway (2021) could not attest to the differences in frequencies of *like* uses among Irish (L1) and Chinese or Polish (L2) speakers of English. Moreover, neither proficiency in English nor length of stay in Ireland was shown to influence the use of this discourse marker (Diskin-Holdway, 2021). A difference between the L1 and L2 speakers, however, was detected in the positioning of *like* within the clause. The former showed higher frequencies of clause-final *like* (Diskin-Holdaway, 2021). Yet one additional crucial finding identified by Diskin-Holdaway (2021) was that the use of *like* as a discourse marker was characterized by a high individual variation. Similarly, Liao (2009) also found a large individual variation among the six Chinese teaching assistants living in the United States. At first sight, this may be interpreted as a gender effect, namely, that, as is sometimes reported, female speakers use discourse markers more frequently than their male peers (Liao, 2009, p. 1321). Yet in Liao's (2009) study, one of the female speakers showed considerably lower rates than the male speakers, underlining that a simple generalization would be imprecise. Furthermore, *like* appeared more frequently in personal interviews than in discussions. Thus, the register turned out to be more conclusive than gender (Liao, 2009, p. 1326). Finally, she remarked that L2 speakers should not be considered as homogenous groups but rather analyzed as individuals with distinct and complex identities (Liao, 2009, p. 1326).

A high level of individual variation was also detected in Rüdiger (2021), even though *like* was overall found to be a prominent feature of the young and educated Korean speakers of English investigated in her study. These findings are based on personal interviews conducted between the author (a young female German speaker of English) and one Korean speaker (i.e., the setting is quite comparable to the setting of the current study, see *The LARES Project*). She identified that, on average, each speaker used approximately eight instances of *like* per one thousand words (Rüdiger, 2021, p. 7).[5] Strikingly, four speakers did not use *like* as a discourse marker at all, whereas one speaker used it as frequently as 70 times per one thousand words (Rüdiger, 2021, p. 8). Even though the speaker variation was quite large, Rüdiger (2021, p. 8) found effects of time spent in an English-speaking country and self-reported proficiency in English but no effect with respect to gender. Those who had spent time abroad and those who reported having higher proficiency in English used *like* more frequently (Rüdiger, 2021, 9).

In summary, the most important variables argued to have an impact on the use of the discourse marker *like* are proficiency in English, input to (naturalistic) English language use, length of stay in an English-speaking country, the status of English (native speaker versus ESL versus EFL), age, and register or the context of language use, including formal versus informal speech. Moreover, a special role may be assigned to influence from American English, and the L1 of the English learners might also have an influence on the use of *like* as a discourse marker. Finally, the effect of gender remains inconclusive, as some studies

---

[5]Rüdiger (2021) used the same classification as D'Arcy (2017). This means that she differentiated between discourse marker and discourse particle uses. The frequency reported here represents the sum of both uses.

acknowledge the role of gender (with higher frequencies for female speakers), whereas others do not find differences between female and male speakers.

Based on the preceding discussion on the use of the discourse marker *like* among different (mainly L2) users of English with at times conflicting results, the current study seeks to answer the following two research questions, shifting the focus from L2 English to ELF users.

RQ1: How frequent is the discourse marker *like* in university student oral interviews conducted in Sharjah and how does its use differ in comparison to other varieties of English?

RQ2: What is the influence of social variables (e.g., gender, citizenship, and year of birth) on the frequency of use?

Before answering these research questions by presenting the results in *Results*, the next section will outline the methodology of the study.

METHODOLOGY

The following three subsections 1) briefly describe the larger project this study is part of, 2) present the participants as well as the social variables relevant to the analysis, and 3) introduce the spoken corpus as well as explain the coding and the subsequent corpus analysis.

The LARES Project

The study employs a subsample of a spoken corpus that consists of semi-structured interviews, approximately 30 min each. In total, 116 students attending the American University of Sharjah participated in the interviews, out of which 58 randomly selected interviews (50%) make up the corpus used in the current study. The participants come from a variety of linguistic backgrounds and include both Emirati and non-Emirati populations. As indicated earlier, some of the interviewees live in Sharjah and others in neighboring emirates, for instance, Dubai. The interviews were conducted by three young, female researchers in March and April 2019 as part of a larger project on LARES (2019–2021). The author of this study was one of the interviewers. Two of the interviewers have a German background, and one has an Iranian background but grew up in Germany. Each interview included questions targeting family background, educational history, specific language biographies, and attitudes towards English, Arabic, and other languages in the students' repertoires. Prior to participating in the interviews, the students completed a comprehensive online survey that was specifically developed for the project LARES based on Siemund et al. (2014) and Leimgruber et al. (2018).[6] The

TABLE 1 | Overview of participants (citizenship and L1).

Citizenship	L1 Arabic	L1 English	Other L1	Total
Arab expatriate	12	8	—	20
Emirati	6	6	—	12
Other	—	4	3	7
South Asian	—	18	1	19
Total	18	36	4	58

survey consisted of questions and agreement statements concerning the demographic, educational, and socioeconomic background, as well as language use and language attitudes (see Siemund et al., 2020; Siemund and Leimgruber, 2021).[7] With this, it is possible to outline and assess the migration history, educational background, language history, and attitudes towards English and Arabic. This detailed information nicely complements the spoken corpus and does not only allow investigating the use of *like* but makes it possible to correlate it with different social (non-linguistic) variables.

A limitation of the dataset used in the current study is that only university students' language production is included. This necessarily restricts any claims made further below to this particular population. Moreover, only one specific genre (i.e., one-to-one interviews) of one specific English variety (English spoken in the UAE) at one specific point in time (2019) is considered here. Ideally, including "a wide array of genres" of several ELF varieties, potentially even at different times or with different generations, would be a broader basis for this kind of investigation (Laitinen, 2020, p. 430). As such, it is not possible to provide a language change perspective per se. Instead, this study can only document the use of *like* from a synchronic perspective and in relation to other studies.

LARES Participants and Social Variables

The current corpus includes 58 semi-structured interviews, conducted with female ($n = 27$) and male ($n = 31$) university students. Their mean age is 20.2 (SD = 1.5) and ranges from 17 to 24. Fifteen different citizenships are represented in this sample, grouped into four distinct groups, namely, Emirati, Arab expatriate, South Asian, and other.[8] The students are part of all four colleges at the American University of Sharjah, that is, Architecture, Art, and Design ($n = 3$); Arts and Sciences ($n = 14$); Engineering ($n = 28$); and business administration ($n = 13$).

In the online survey, the students were asked to rank the languages they speak, starting with the language they are most proficient in. Thirty-six students ranked English first, followed by 18 who ranked Arabic first. Four students reported another language to be their most proficient or dominant language.

---

[6]As part of LARES, 692 students completed the online questionnaire. Part of the questionnaire was to signal a willingness to additionally participate in a semi-structured interview. From those indicating their interest, 116 students finally participated in the semi-structured interview. These interviews make up the LARES corpus.

[7]The sociolinguistic variables applied in this study reflect Anglo-American/ European settings, i.e., gender, citizenship, and language background. By choosing these variables, comparability with previous research and other sociolinguistics studies can be assured.

[8]The 15 citizenships are Argentina, Bahrain, Bangladesh, Egypt, India, Iraq, Jordan, Lebanon, Morocco, Pakistan, Russia, Sri Lanka, Tunisia, the UAE, and the USA.

TABLE 2 | Number of participants per number of languages.

Number of languages	Total
2	9
3	20
4	19
5	7
6	3

The language ranked the highest will henceforth be referred to as L1. Note that this does not necessarily have to be the native language or the language acquired chronologically first. This is visible from **Table 1**, where citizenship and the respective L1 are presented. In the online survey, 14 of the Emiratis or Arab expatriates indicated English to be their strongest language (L1), yet in the interviews, they reported Arabic to be their native language.

In both the online survey and during the interviews, the students were asked how many languages they knew or had some proficiency in. Particularly during the interviews, the interviewers stressed that all languages counted, i.e., also those in which the participants had a relatively low proficiency such as a foreign language studied for some years in school a while ago. When comparing the responses from the survey with those of the interviews, it is quite striking that 27 students reported knowing a higher number of languages during the interviews. At times, they remembered halfway through the conversation that they had learned French in school for some years, for instance. Here, the numbers from the interviews are used, as they seem to better reflect the multilingual repertoires of the students. The majority of the current sample has at least some proficiency in either three or four languages (*n* = 39). The overall distribution can be found in **Table 2**.

In the online survey, the students were asked to self-assess their proficiency in English. Using a scale from one (mastery) to six (beginner), they should assign a score to listening/understanding, speaking fluency, reading proficiency, and writing proficiency separately. The proficiency score used in the current study is the resulting mean of the four individual measures.

In addition to these rather traditional sociolinguistics variables, a new variable was created. This variable is based on the interview data and is called "English usage score." The main motivation for this analysis step was that, as will become apparent in the *Results* section, the previously described social variables turned out not to be significant predictor variables of the distribution of the discourse marker *like*. Moreover, previous research has shown that learners who are exposed to "naturalistic English" seem to use the discourse marker *like* more frequently (Gilquin, 2016, p. 244). Since all participants are students enrolled in an American University, hence, the medium of instruction is English, and they are all clearly exposed to English on a regular basis. However, there may be differences in English usage outside of university as well as stemming from their school education. For instance, the use of English (social) media may show dissimilarities. In addition, some students may have attended

schools with English as the educational language, whereas others may have attended schools using other instructional languages such as Arabic.

In order to get a better understanding, each interview was coded for seven sub-variables, which are "English TV/movies," "English-speaking country," "Length of stay in an English-speaking country," "Language of instruction in school," "School system," "Pick one language to keep,"[9] and "Best language." After coding, two variables were dropped, namely, "English TV/movies" and "Pick one language to keep." The former was removed, as practically everyone reported watching movies or TV in English, and more specific details (such as "frequently" or "only occasionally") were impossible to determine from the interviewees' responses. Two students did not specifically mention that they watched films or TV in English. Yet they did not deny it either but rather stated that they enjoyed Hindi and Arabic movies a lot. Thus, it can be assumed that all students are in some way or another exposed to English *via* movies or TV, and this variable could therefore be discarded. Unfortunately, in ten interviews, the question as to which language they would keep if they had to choose one was not asked. To avoid the reduction of the dataset, this variable was also removed from the English usage score. Finally, four of the remaining sub-variables were coded as two, as they belonged to the same category (see the explanation of the final variables below).

The refinement of the English usage score is thus composed of three different measures, "English-speaking country" (i.e., number of times traveled to an English-speaking country: 0, never; 1, once or twice/infrequently; 2, regularly, frequently, often), "Language of instructions" (i.e., school education received in: 0, Arabic or another language than English; 1, English and another language; 2, English only), and "Best language" (i.e., the language they feel most comfortable with: 0, Arabic or another language than English; 1, English and another language; 2, English). The resulting English usage score is the sum of the three measures ranging from zero to six.

LARES Corpus: Data Coding and Analysis

The recordings of 58 interviews that represent 50% of the entire dataset were transcribed by one person and checked by another one. XML tags were used to distinguish between the interviewee's (iwe) and interviewer's (iwr) utterances. For the current study, exclusively, the interviewee data were analyzed. This part of the corpus contains 139,630-word tokens of orthographically transcribed speech, with an average of 2,407-word tokens per interview file. Strikingly, *like* turned out to be the third most frequent word form (*n* = 3,937), which clearly supports the importance of investigating its usage.

All instances of *like* were extracted from the LARES corpus using the concordance program AntConc (Anthony, 2018). In a second step, all hits were manually annotated. In the first round of coding, the discourse marker *like* was differentiated from other

---

[9]One of the questions the interviewees were asked was "If you have to choose only one of the languages you know, which one would it be?"

uses of *like*, such as verbs, comparative prepositions/complementizers, nouns, suffix, quotative *like*, or repetitions, following D'Arcy (2017, pp. 3–13) and Schweinberger (2014, pp. 140–143). In the second round of coding, the initial decision was checked and, if necessary, corrected; and each instance of a discourse marker was further categorized depending on the clausal position. This coding step was based on Schweinberger's (2014, pp. 146–149) coding scheme. Thus, the current study distinguishes between clause-initial (INI), clause-medial (MED), clause-final (FIN), and non-clausal (NON) *like*. Examples (4) to (7) represent each type.

(4) *Like English is easy to communicate but not every local knows English.* (f20) [INI]

(5) *I mean um my academic career is like mostly English* (m3) [MED]

(6) *[. . .] but we took like English as a course like.* (f14) [FIN]

(7) *Okay Arabic is for me it's the like the its more complex than any of the others.* (f22) [NON]

The absolute frequencies of the discourse marker *like* were then, in order to ensure comparability across the interview files, normalized to the basis of 1,000 words. As a next step, each social variable was investigated separately to assess its relation with the frequency of *like*. For this, the variables identified as influencing the use of discourse markers in general or *like* in particular (see *Discourse Marker Like*) and those assessed *via* the online survey (see *Discourse Marker Like* and *LARES Participants and Social Variables*) were used. Following this monofactorial analysis, a generalized linear regression analysis was run.

RESULTS

The following three subsections present the results of the corpus analysis. First, the overall (absolute) frequency of *like* is discussed. Second, the uses of the discourse marker *like* versus social variables are exhibited based on the normalized frequencies, and third, the most frequent *like*-users are examined as a separate cohort.

Frequency of *Like*

Research question 1 asked how frequently the discourse marker *like* appeared in the spoken corpus and how this differed in comparison to other varieties of English. On the whole, *like* is the third most frequently used word in the interviewees' utterances ($n = 3,937$), with 2,951 (75%) uses as a discourse marker and 986 (25%) other uses.[10] The classification of the discourse marker uses shows that clause-initial ($n = 1,466$, 50%) and clause-medial ($n = 1,206$, 41%) make up the largest part, whereas non-clausal ($n = 235$, 8%) and particularly clause-final ($n = 44$, 1%) uses are

relatively infrequent. The mean frequency per 1,000 words (ptw) across the entire corpus is 19.5 (median: 16.0), yet the relatively high SD of 14.75 shows that the individual variation among the speakers is comparably large. The lowest frequency is 0.51 ptw, and the highest is 55.14 ptw. The visualization across 11 intervals (see **Figure 1**) shows that lower frequencies (from 1 to 20 or perhaps even until 30 discourse marker uses ptw) are more often represented in the dataset than higher frequencies (from 30 to 55 ptw).

A closer look at the social background variables in combination with the normalized frequencies of the discourse marker *like* provides a more detailed picture.

Like Versus Social Variables

To answer the second research question, namely, how the social background of the speakers influences the use of *like*, seven different social variables will be juxtaposed with the normalized frequencies of the discourse marker. These are gender, citizenship, L1, year of birth, number of languages, college, self-assessed proficiency in English, and the English usage score. For all statistical tests, the free software R (R Core Team, 2020) is used.

Even though the mean frequency of the female students (21.1) is slightly higher than the mean frequency of their male peers (18.0), the Wilcoxon test did not return a statistically significant difference ($W = 468.5$, $p = 0.22$). This suggests that there is no difference between male and female speakers with respect to the frequency (ptw) of the discourse marker *like* in the current sample.

A similar result was obtained for the Kruskal–Wallis test when testing citizenship (Emirati, Arab expatriate, and South Asian)[11] versus normalized frequency of discourse marker *like*. No statistical significant difference was attested (H (3) = 2.5, $p = 0.47$), confirmed by nonsignificant post-hoc tests.

An equally nonsignificant difference resulted from the Wilcoxon test ($W = 376.5$, $p = 0.17$) when comparing the means of the discourse marker use with respect to L1 English and Arabic. This means that citizenship is also not a good predictor when explaining frequency differences in the current sample.

When correlating the age of the participants (range from 17 to 24) with the normalized frequencies of *like*, a very small negative correlation coefficient was obtained ($r = -0.09$); yet this very weak negative correlation did not reach statistical significance ($p = 0.53$). Perhaps the age range of the participants is too small to find a correlation. A less homogeneous group may indeed return a significant correlation of age and frequency of *like*.

---

[10]The pronouns *I* and *it* are the two most frequent words in the LARES corpus, followed by *like*, *in*, *the*, and *to*.

[11]For the monofactorial statistical analysis, only Emirati, Arab expatriate, and South Asian students were considered. Those with another citizenship had to be excluded, as this group consisted of four students only. The same procedure was followed for L1 as well as college. Those who reported another L1 than English or Arabic were not considered for the statistical analysis, because only four students belonged to this group. The college of Architecture, Art, and Design ($n = 3$) was excluded, and only the colleges of arts and sciences, business administration, and engineering were featured in the statistical analysis.
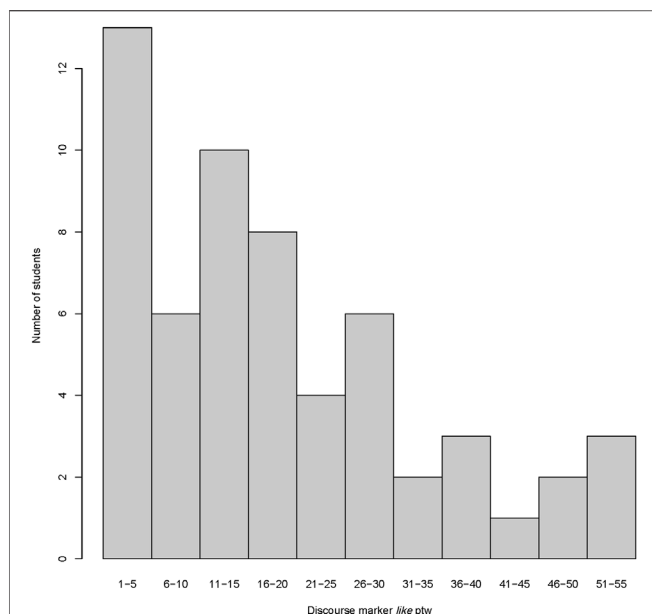
FIGURE 1 | Distribution of discourse marker *like* ptw across normalized frequency intervals.

The number of languages in the students' repertoires was also correlated with the normalized frequencies of the discourse marker *like*. The correlation coefficient is positive even if very small (0.11), and it does again not reach statistical significance ($p = 0.20$). Once more, no statistically significant difference of a background variable, here the number of languages, can be attested.

The Kruskal–Wallis test was used to compare the mean frequencies of *like* among the students attending the different colleges including the Colleges of Arts and Sciences, Business Administration, and Engineering. The results show that there is no statistically significant difference across the three colleges based on the Kruskal–Wallis test statistics ($H (2) = 2.70$, $p = 0.26$) as well as following post-hoc tests.

A correlation analysis of the self-assessed English proficiency versus the normalized frequencies of *like* returned a very small negative correlation ($r = -0.04$), which is not statistically significant ($p = 0.77$). The following Wilcoxon test based on two groups (low versus high proficiency) derived *via* the median split (1.625) does not return a statistically significant difference between the two groups ($W = 408.5$, $p = 0.22$) either.

Finally—and as initially indicated because all preceding variables returned no statistically significant differences—the English usage score was correlated with the normalized frequencies of the discourse marker *like*. The resulting correlation coefficient is once again very small ($r = -0.11$), and this negative correlation is also not statistically significant ($p = 0.42$). As before, a Wilcoxon test comparing those who have a low English usage score (0–3) versus those with a higher score (4–6) does not reach statistical significance either ($W = 342$, $p = 0.20$).

What this means is that none of the social variables considered in this study turn out to statistically significantly differ with respect to the use of the discourse marker *like*. This was ultimately confirmed with a generalized linear regression analysis, using a Poisson regression. The social variables did not statistically significantly contribute to explaining the variance in the frequency of the discourse marker *like*.[12] A logical next step is then to have a closer look at those students who have the highest ratios of *like* used as a discourse marker and to investigate if these students share specific characteristics that could not be identified with the preceding analyses.

Most Frequent *Like*-Users

As an extension of research question 2, the ten most frequent *like* users were separately analyzed in order to identify features they share. This additional analysis step may allow further conclusions as to which social variables are particularly strongly associated with high frequencies of *like*. The ten speakers with the highest ratios of *like* (between 33 and 55 ptw) form a relatively heterogeneous group. A close inspection of the characteristics they have in common reveals that they present the entire range of all possible variable manifestations. Five male and female speakers are represented, some reported to know only two or three languages, others know four or more, and there are Emirati, Arab expatriate, and South Asian students present. The only citizenship group missing is other, but we have to keep in mind that in the entire sample, there were only four students with citizenship other than the three mentioned here. One interesting indication may perhaps be the college associated with the students. All but one attend the college of engineering (one is part of the college of arts and sciences). Yet it has to be acknowledged that nearly 50% of all students are part of engineering, which clearly increases the likelihood of appearing among this subsample as well and may thus be a sampling condition instead of a finding. Moreover, more of these ten students reported English to be their L1 ($n = 6$), and fewer ranked Arabic first ($n = 4$). With this ratio of 3:2, there are fewer English L1 speakers than in the overall sample (which has a ratio of 2:1). Half of the participants indicated that they have very high competencies in English (a self-assessed proficiency score of 1.25 or lower), whereas the other half rated their English skills slightly lower (four scored two or lower, and one scored 3.75). The mean score, however, is only marginally, if at all, higher among the ten students (1.65) in comparison to the entire sample (1.69), and it is in general relatively high. Even though so far self-assessed proficiency in Arabic has not played a role, it had been considered for this final analysis. The scoring procedure was the same as for English (i.e., from 1 to 6), and among the ten students, there are some with high skills in Arabic (1) up to relatively low skills (3.75), in addition to two students who indicated that they do not speak any Arabic.

---

[12]Two separate generalized linear regressions were fit. The first included all main effects, and the second allowed two-way interactions of all variables. *Via* model selection (backward stepwise) based on *p*-values (threshold 0.05) (see, for example, Gries, 2021, p. 366), all predictors (interactions as well as main effects) had to be dropped, as they did not contribute statistically significantly to the models. No predictors remained.

All in all, no specific pattern could be identified, which at first may seem even disappointing. Yet, and this will be argued for in the next sections, this may indeed produce some interesting implications. Arguably, there is quite a bit of variation across the students, yet none of the variables considered seem to have a particularly strong association with the use of the discourse marker *like*. In the following discussion, these results will be looked at in relation to findings from other studies investigating this discourse marker.

DISCUSSION

The results indicate that *like* used as a discourse marker is a prominent feature among the students investigated here. The first subsection of the discussion considers the overall frequency of *like* in the semi-structured interviews and discusses this in light of the earlier studies investigating different varieties and different speakers of English. The second subsection looks at *like* in combination with the social background variables of the students.

Frequency of *Like*

The overall frequency of the discourse marker *like* in the current study is surprisingly high (mean = 19.50 ptw; SD = 14.75), particularly so when compared to the frequencies found in Schweinberger (2014). He reported mean frequencies for British, Indian, New Zealand, Philippine, and Canadian English ranging from 0.45 to 4.39 instances of *like* per one thousand words (Schweinberger, 2014, p. 185). This may be partly due to the specific genre used in the current study, namely, semi-structured, personal, and relatively informal interviews, as opposed to Schweinberger (2014) who relied on the International Corpus of English (ICE), which represents various types of spoken language. This genre-related argument can be supported with the results discussed in Fuller (2003) investigating American English. She noticed that *like* appeared relatively frequently in her study (11.6 ptw). These findings were also based on interviews. In interviews, the use of discourse markers may be pragmatically useful, particularly in interviews perceived as relatively informal and personal, even though Fuller (2003) thought it was remarkable that the speakers in her study used *like* with such a high frequency. The reason for her surprise was the presumed stigmatization of *like* and its association with "a lack of intelligence" (Fuller, 2003, p. 369). This, however, could not be confirmed in Fuller's (2003) research with young speakers of US English. Yet the frequencies reported in her study are still considerably lower compared to the mean frequency found in the LARES corpus.

In a study equally based on personal one-to-one interviews, Rüdiger's (2021) investigation of Korean English found a mean frequency of *like* used as a discourse marker of approximately eight per one thousand words. However, similar to the current study, she noticed high rates of internal variation with one speaker having a frequency of 70 uses of *like* per one thousand words. This is even more extreme than what was found in the LARES corpus, where the most frequent *like*-user had a frequency of 55.14 uses per one thousand words.

Two potential explanations may be feasible to interpret the high speaker variability as well as the overall high frequency of *like*. On the one hand, ELF may be particularly prone to show variability across speakers of one speech community, irrespective of their linguistic or social background (see *Like Versus Social Variables* for more details). On the other hand, from a language change perspective, such a generally high frequency of *like* may also be explained with the specific setting in which English is used. Users of ELF have a multilingual background and comparable multilingual contexts have been shown to accelerate language change (see, for example, Laitinen, 2020, p. 428). The use of *like* as a discourse marker has seen a recent increase in varieties of English (D'Arcy, 2017, pp. 14–15), and the multilingual context present in the UAE may be responsible for an even greater frequency increase. In addition, influence through (digital) media, particularly from the United States, may further advance this development (for more information about the influence of the United States, see the following section).

Like Versus Social Variables

The most important finding is that the social background of the speakers cannot explain the variability identified in the use of the discourse maker *like*. The statistical analysis presented in *Like Versus Social Variables* demonstrated that the female students did not use the discourse marker *like* significantly more frequently than their male peers, even though the mean frequency of the female students (21.1 ptw) was slightly higher than that of the male students (18.0 ptw). Both Fuller (2003) and D'Arcy (2007) identified a female speaker lead among speakers of American English, yet Schweinberger (2014, p. 393) presented a more diverse picture and argued for "variety-specific" differences with respect to gender. Moreover, Rüdiger (2021) could not identify differences with respect to gender either. Perhaps, in our modern and globalized era, it may be time to focus more on other variables than on the binary variable gender or to approach it as less binary. Cleary, it is a very convenient variable, relatively straightforward to code and include in research. Yet it may not necessarily be a reliable indicator of discourse marker usage (or other markers, for that matter). Perhaps a more fine-grained or scalar category, potentially not with respect to gender but perhaps a personality marker instead (i.e., in relation to the big five personality traits), would shed more light on uses of *like* as a discourse marker.

The slight tendency that younger students used the discourse marker *like* more often than older students would further substantiate that the use of *like* is increasing in the English of the UAE students investigated here. However, this needs to be taken with great caution, as the negative correlation was extremely small and did not reach statistical significance. This may be due to the small age range of the students and further research including older (and also younger) groups of people may further substantiate this observation.

Furthermore, this study did not identify statistically significant differences in the frequency of *like* pertaining to

citizenship, L1, number of languages, or college. A few words pertaining to citizenship are in order. Similar to gender, citizenship as a definite category may not necessarily reflect the multilingual and multicultural identities of these students. Perhaps the participants could be better classified as global citizens having a "modern global identity" (Fuller, 2020, p. 167). Many of them have moved multiple times across countries or even continents, and they converse on a regular basis with many different individuals coming from diverse backgrounds. Yet a relative closeness towards especially American English but also British English cannot be denied. Many of the students have either followed an American or a British school curriculum, they all attend an American university, and quite a number of students associate their English either with American or British English or aim at using one of the two varieties. Particularly because of their affiliation with an American university, influence from the United States could be argued to be relatively prominent, which, in turn, could have an effect on the use of the discourse marker *like* as shown by Müller (2005). Nevertheless, parallels may be drawn to what Fuller (2020, p. 167) means when she talks about "a modern global identity that is not linked to any particular nationality" where English is "used to convey a cosmopolitan connotation." Even though Fuller (2020) argued this to be true for English in Germany—and this context may be understood as distinctly different from the UAE context—it is still worthwhile extending this to the current study.

Thus, in light of the literature review introduced above, it should be acknowledged that L2 learners are not straightforwardly comparable to the ELF users considered in the current study. The main reason is that the LARES participants are advanced speakers of English who use this language during their studies and mostly also outside of the university. Hence, proficiency, which had been identified as a predictor variable of usage rates of *like* (see Hasselgren, 2002; Neary-Sundquist, 2014), is necessarily high among all interviewees (which was already visible in the self-reported proficiency ratings). It may thus be less surprising that the variable self-assessed proficiency did not turn out as a significant predictor. Including further speakers with more variability in relation to proficiency in English may further corroborate the claim that proficiency impacts the use of the discourse marker *like*. Moreover, as indicated in Gilquin (2016), it is access to naturalistic language input outside of the English language classroom that particularly boosts the acquisition and use of discourse markers (see also Liao, 2009, p. 1314). More than half of the participants reported that English was their dominant language. Even the remaining students, whose dominant language is either Arabic or another language, can be assumed to have access to English on a regular basis, first because of their studies (English is the exclusive medium of instruction) in addition to media consumption and interaction with peers and partly even within the home. English truly plays an important role in the lives of these young students. They are

in fact confident speakers and may also identify with this language. The following two quotes, taken from the LARES interviews, underline this.

(8) *So like I said earlier when it comes to like English it's very it's very much like a tool.* (m28)

(9) Answer to the questions of which language to keep, if only one could be kept: *English.* [Interviewer: *Why*?] *Coz like, the other languages* [English, Hindi, Urdu, Arabic, Farsi, Russian] *are like small parts of my life, but like my life runs in English.* (m12)

English is understood as a tool, it dominates the lives of the students, and it may even replace the native language in terms of daily use and importance. Employment aspirations and the prospective importance of English with respect to future careers are certainly two of the driving factors. Nevertheless, it is imperative to acknowledge that other languages, particularly Arabic, also play a role (Thomas, 2021). Yet as student m12 admits in (9), these languages compete on a different level with English and typically are of secondary importance. To be clear, this is true for the particular population considered here and may not necessarily be generalized to other groups residing in the UAE (or even students attending another university).

Moreover, what has to be acknowledged is that the discourse marker use of *like* seems to be a relatively prominent feature of the English repertoires of the UAE students. In a sense, these speakers could therefore be seen to be somewhat comparable to the Korean English speakers analyzed in Rüdiger (2021). Clearly, Korea and the UAE are two quite different geographical locations and social realities, but perhaps the younger generations in a globalized world are not that different anymore. As argued above, young people and their access to media may make it relevant to assess the global context in addition to or instead of simply regarding citizenship or country of residence. However, the higher rates of *like* among the UAE students in comparison to the Korean speakers in Rüdiger's (2021) study may indeed hint at higher English proficiency and more frequent English use of the former.

Finally, even though many speakers considered here made use of *like* as a discourse marker relatively frequently, this was not true for all speakers. The observed individual variation might be the result of the particular multilingual setting investigated. One characteristic of ELF encounters is that speakers with various linguistic backgrounds are in contact. Mauranen describes this context as a speech situation "[w]ith vast numbers of similects coming into contact with each other" (2017, p. 230). What this means is that different Englishes are in contact, which itself are influenced by the L1 or L1s of each person (at least in the case of non-native speakers of English) in addition to the specific English acquisition setting (Mauranen, 2017, p. 227). Because of this heterogeneity, Mauranen (2017, p. 230) assumes a substantial amount of variability in ELF uses. The present study provides evidence for the high variability of *like* among highly proficient

English users in the UAE. Whether this holds true for other features of the English spoken in the UAE remains to be seen in future research.

CONCLUSION AND OUTLOOK

This study contributes to the understanding of the use of the discourse marker *like* in ELF, more precisely in the English spoken by university students in the UAE. This is important for world Englishes at large, as ELF has so far not been in focus with respect to the use of discourse markers.

Like used as a discourse marker appeared to be a frequent and prominent feature in the semi-structured, personal, and relatively informal interviews between young female researchers and young multilingual and multicultural university students in Sharjah. The social variables included in this study (i.e., gender, citizenship, L1, year of birth, number of languages, college, self-assessed proficiency in English, and English usage score) returned no statistically significant differences. Even though variability among the students was relatively high, all participants used the discourse marker *like* at least once. Therefore, it could be argued that *like* presents a prominent marker of the English spoken by this elite and educated group of speakers. The relatively high frequency of *like* might support the hypothesis of accelerated language change in multilingual settings such as the UAE, where different Englishes of speakers with various L1s are in contact. Moreover, high individual variability among the heterogeneous speakers might be an additional characteristic of ELF users.

What this study cannot provide is a broader view of other groups living in the UAE. Future studies focusing on speakers other than university students may find out how *like* is distributed more generally in the English spoken in the UAE. This study is thus only a starting point for understanding the English used in the UAE, and further research, including studies on other linguistic features, is urgently needed.

Finally, Siemund (2022) argued that even though English in the UAE can be considered a lingua franca, it may in fact rather be categorized as a second language. The findings of the current study may perhaps offer modest support of this claim, particularly so for young university students,

demonstrated by the high rates or frequent use of *like* as a discourse marker.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of the agreement signed in the IRB form. Requests to access the datasets should be directed to EL, eliane.lorenz@anglistik.uni-giessen.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Office of Research, American University of Sharjah, UAE, *via* the IRB Application Form. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author was involved in the data collection and confirms being the sole contributor of this article.

ACKNOWLEDGMENTS

REFERENCES

Ahmad, R. (2016). Expatriate Languages in Kuwait: Tension between Public and Private Domains. *J. Arabian Stud.* 6 (1), 29–52. doi:10.1080/21534764.2016.1192767

Ahmed, K. (2021). "Linguistic and Semiotic Landscapes of Dubai," in *The Multilingual Global Cities: Singapore, Hong Kong, Dubai*. Editors P. Siemund and J. R. E. Leimgruber (Singapore: Routledge), 185–202.

Al-Issa, A. (2021). "Multilingualism, Language Management, and Social Diversity in the United Arab Emirates," in *Multilingual Global Cities: Singapore, Hong Kong, Dubai*. Editors P. Siemund and J. R. E. Leimgruber (Singapore: Routledge), 116–130.

Anthony, L. (2018). *AntConc (Version 3.4.4.0) [Computer Software]*. Tokyo, Japan: Waseda.

Boyle, R. (2012). Language Contact in the United Arab Emirates. *World Englishes* 31 (3), 312–330. doi:10.1111/j.1467-971x.2012.01749.x

Boyle, R. (2011). Patterns of Change in English as a Lingua Franca in the UAE. *Int. J. Appl. Linguistics* 21 (2), 143–161. doi:10.1111/j.1473-4192.2010.00262.x

Cook, W. R. A. (2017). More Vision Than Renaissance: Arabic as a Language of Science in the UAE. *Lang. Pol.* 16, 385–406. doi:10.1007/s10993-016-9413-3

Corrigan, K. P., and Diskin, C. (2019). 'Northmen, Southmen, Comrades All'? the Adoption of Discourse like by Migrants north and South of the Irish Border. *Lang. Soc.* 49, 745–773. doi:10.1017/s0047404519000800

D'Arcy, A. (2017). *Discourse-Pragmatic Variation in ContextEight Hundred Years of LIKE*. Amsterdam: Benjamins.

D'Arcy, A. (2007). Like and Language Ideology: Disentangling Fact from Fiction. *Am. Speech* 82, 386–419.

Davidson, C. M. (2005). *The United Arab Emirates: A Study in Survival*. Boulder, CO: Lynne Rienner.

Diskin, C. (2017). The Use of the Discourse-Pragmatic Marker 'like' by Native and Non-native Speakers of English in Ireland. *J. Pragmatics* 120, 144–157. doi:10.1016/j.pragma.2017.08.004

Diskin-Holdaway, C. (2021). You Know and like Among Migrants in Ireland and Australia. *World Englishes*. doi:10.1111/weng.12541

Dubai Population (2020). Dubai Population 2020. Available at: https://worldpopulationreview.com/world-cities/dubai-population (Accessed September 1, 2020).

Filppula, M., Klemola, J., Mauranen, A., and Vetchinnikova, S. (2017). "Changing English: Global and Local Perspectives," in *Changing English. Global and Local Perspectives*. Editors M. Filppula, J. Klemola, A. Mauranen, and S. Vetchinnikova (Berlin/Boston: DeGruyter Mouton), xi–xiii. doi:10.1515/9783110429657-204

Fuller, J. M. (2020). "English in the German-Speaking World: Immigration and Integration," in *English in the German-Speaking World*. Editor R. Hickey (Cambridge: Cambridge University Press), 165–184.

Fuller, J. M. (2003). Use of the Discourse Marker like in Interviews. *J. Sociolinguistics* 7 (3), 365–377. doi:10.1111/1467-9481.00229

Fussell, B. (2011). The Local Flavour of English in the Gulf. *English Today* 27 (4), 26–32. doi:10.1017/s0266078411000502

Gabrys, M. (2017). "Like as a Discourse Marker in Different Varieties of English. A Contrastive Corpus-Based Study,". Master's Thesis (Louvain: Université catholique de Louvain). Available at: https://dial.uclouvain.be/memoire/ucl/fr/object/thesis%3A11925/datastream/PDF_01/view (Accessed January 20, 2021).

Gilquin, G. (2016). "Discourse Markers in L2 English," in *New Approaches to English Linguistics: Building Bridges*. Editors O. Timofeeva, A.-C. Gardner, A. Honkapohja, and S. Chevalier (Amsterdam: Benjamins), 213–249. doi:10.1075/slcs.177.09gil

Government of Dubai (2019). Population and Vital Statistics. Available at: https://www.dsc.gov.ae/en-us/Themes/Pages/Population-and-Vital-Statistics.aspx?Theme=42 (Accessed September 1, 2020).

Gries, S. Th. (2021). *Statistics for Linguistics with R. A Practical Introduction*. 3rd Ed. Berlin: De Gruyter Mouton.

Hasselgren, A. (2002). "Learner Corpora and Language Testing," in *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Editors S. Granger, J. Hung, and S. Petch-Tyson (Amsterdam: Benjamins), 143–173. doi:10.1075/lllt.6.11has

Hopkyns, S. (2017). "A Conflict of Desires: Global English and its Effects on Cultural Identity in the United Arab Emirates,". Dissertation (Leicester: University of Leicester).

Hopkyns, S. (2021). "Multilingualism and Linguistic Hybridity in Dubai," in *Multilingual Global Cities: Singapore, Hong Kong, Dubai*. Editors P. Siemund and J. R. E. Leimgruber (Singapore: Routledge), 248–264.

Laitinen, M. (2020). Empirical Perspectives on English as a Lingua Franca (ELF) Grammar. *World Englishes* 39 (3), 427–442. doi:10.1111/weng.12482

Leimgruber, J. R. E., Siemund, P., and Terassa, L. (2018). Singaporean Students' Language Repertoires and Attitudes Revisited. *World Englishes* 37 (2), 282–306. doi:10.1111/weng.12292

Leimgruber, J. R. E., and Siemund, P. (2021). "The Multilingual Ecologies of Singapore, Hong Kong, and Dubai," in *Multilingual Global Cities: Singapore, Hong Kong, Dubai*. Editors P. Siemund and J. R. E. Leimgruber (Singapore: Routledge), 1–15.

Leuckert, S., and Rüdiger, S. (2021). Discourse Markers and World Englishes. *World Englishes*. doi:10.1111/weng.12535

Liao, S. (2009). Variation in the Use of Discourse Markers by Chinese Teaching Assistants in the US. *J. Pragmatics* 41, 1313–1328. doi:10.1016/j.pragma.2008.09.026

Madichie, N. O., and Madichie, L. (2013). City Brand Challenge 101: Sharjah in a Globalised UAE Context. *Ijbg* 11 (1), 63–85. doi:10.1504/ijbg.2013.055316

Mair, C. (2021). "World Englishes: From Methodological Nationalism to a Global Perspective," in *The Bloomsbury Handbook of World Englishes. Volume 1*. Editors B. Schneider and T. Heyd (London: ParadigmsBloomsbury), 27–45.

Mauranen, A. (2017). "A Glimpse of ELF," in *Changing English. Global and Local Perspectives*. Editors M. Filppula, J. Klemola, A. Mauranen, and S. Vetchinnikova (Berlin/Boston: DeGruyter Mouton), 223–254. doi:10.1515/9783110429657-013

Mauranen, A. (2012). *Exploring ELF. Academic English Shaped by Non-native Speakers*. Cambridge: Cambridge University Press.

Mukherjee, J., and Rohrbach, J.-M. (2006). "Rethinking Applied Corpus Linguistics from a Language-Pedagogical Perspective: New Departures in Learner Corpus Research," in *Planing, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop*. Editors B. Kettemann and G. Marko (Frankfurt am Main: Peter Lang), 205–232.

Müller, S. (2005). *Discourse Markers in Native and Non-native English Discourse*. Amsterdam: John Benjamins.

Neary-Sundquist, C. (2014). The Use of Pragmatic Markers across Proficiency Levels in Second Language Speech. *Ssllt* 4 (4), 637–663. doi:10.14746/ssllt.2014.4.4.4

O'Neill, G. T. (2014). Just a Natural Move towards English: Gulf Youth Attitudes towards Arabic and English Literacy. *Learn. Teach. Higher Educ. Gulf Perspect.* 11 (1), 1–21.

Pacione, M. (2005). Dubai. *Cities* 22 (3), 255–265. doi:10.1016/j.cities.2005.02.001

Parra-Guinaldo, V., and Lanteigne, B. (2021). "Morpho-syntactic Features of Transactional ELF in Dubai/Sharjah," in *Multilingual Global Cities: Singapore, Hong Kong, Dubai*. Editors P. Siemund and J. R. E. Leimgruber (Singapore: Routledge), 303–320.

Piller, I. (2017). "Dubai: Language in the Ethnocratic, Corporate and Mobile City," in *Urban Sociolinguistics: The City as a Linguistic Process and Experience*. Editors D. Smakman and P. Heinrich (Abingdon: Routledge), 77–94. doi:10.4324/9781315514659-7

P. Siemund and J. R. E. Leimgruber (Editors) (2021). *Multilingual Global Cities: Singapore, Hong Kong* (Dubai. Singapore: Routledge).

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Randall, M., and Samimi, M. A. (2010). The Status of English in Dubai. *English Today* 26 (1), 43–50. doi:10.1017/s0266078409990617

Rüdiger, S. (2021). Like in Korean English Speech. *World Englishes* 40, 548–561. doi:10.1111/weng.12540

Schneider, E. W. (2007). *Postcolonial English. Varieties Around the World*. Cambridge: Cambridge University Press.

Schweinberger, M. (2014). "The Discourse Marker LIKE: A Corpus-Based Analysis of Selected Varieties of English,". Dissertation (Hamburg: University of Hamburg).

Sharjah Population (2020). Sharjah Population 2020. Available at: https://worldpopulationreview.com/world-cities/sharjah-population (Accessed September 1, 2020).

Siemund, P., Al-Issa, A., and Al-Issa, J. R. E. (2020). Multilingualism and the Role of English in the United Arab Emirates. *World Englishes* 40, 191–204. doi:10.1111/weng.12507

Siemund, P., Al-Issa, A., Rahbari, S., and Leimgruber, J. R. E. (2021). "Multilingualism and the Role of English in the United Arab Emirates, with Views from Singapore and Hong Kong," in *Research Developments in World Englishes*. Editor A. Onysko (London: Bloomsbury), 95–120. doi:10.5040/9781350167087.ch-006

Siemund, P. (2022). *Multilingual Development: English in a Global Context*. Cambridge: Cambridge University Press.

Siemund, P., Schulz, M. E., and Schweinberger, M. (2014). Studying the Linguistic Ecology of Singapore: A Comparison of College and University Students. *World Englishes* 33 (3), 340–362. doi:10.1111/weng.12094

Theodoropoulou, I. (2021). "Socio-historical Multilingualism and Language Policies in Dubai," in *Multilingual Global Cities: Singapore, Hong Kong, Dubai*. Editors P. Siemund and J. R. E. Leimgruber (Singapore: Routledge), 36–80.

Thomas, S. (2021). English in the United Arab Emirates. *World Englishes*. doi:10.1111/weng.12560

Thomas, S. (2016). "The Case of the 'Innocuous' Middle-Class Migrant Employee: English Language Use and Attitudes in Dubai, United Arab Emirates,". Dissertation (West Lafayette (IN): Purdue University).

Vertovec, S. (2007). Super-diversity and its Implications. *Ethnic Racial Stud.* 30 (6), 1024–1054. doi:10.1080/01419870701599465

Wolk, C., Götz, S., and Jäschke, K. (2021). Possibilities and Drawbacks of Using an Online Application for Semi-automatic Corpus Analysis to Investigate

Discourse Markers and Alternative Fluency Variables. *Corpus Pragmatics* 5, 7–36. doi:10.1007/s41701-019-00072-x

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Communication

English from Scratch: Preadolescents' Developing Use of English Lexical Resources in Belgian Dutch

Melissa Schuring* and Eline Zenner

QLVL, Faculty of Arts, KU Leuven, Brussels, Belgium

Working within the framework of the socio-pragmatic turn in anglicism research, this paper adds a developmental sociolinguistic perspective in investigating preadolescents' use of English lexical resources in Belgian Dutch. The so far largely undocumented role of English in the linguistic transition from childhood to adolescence is analyzed through a fieldwork corpus of 15,465 utterances, collected during sociolinguistic interviews with 26 (12 boys, 14 girls) Belgian Dutch preadolescent (6–13 years/o) respondents from a local hockey club. All English lexical material in the corpus was identified and categorized following a three-step identification protocol. This protocol introduces a distinction between recognizable unavoidable English (RUE) and recognizable avoidable English (RAE). Results reveal that, overall, 9.7% of the utterances contain recognizable English (RUE + RAE), with RUE being significantly more frequent than RAE. Our findings further indicate only limited stratification according to traditional socio-demographic parameters and display a number of outliers in the respondent profiles. Closer inspection of these outliers allows the conclusion that in the community of practice studied, English is an emerging youth language marker, typically used when talking about gaming or girl-oriented activities. In sum, we conclude that preadolescents in our sample instrumentalize English for incipient identity work, both on the micro-level (being a gamer, a soon-to-be teenage girl) as on the macro-level (through ingroup and outgroup marking).

Keywords: anglicisms, youth language, developmental sociolinguistics, sociolinguistic interview, preadolescence, Belgian Dutch

1 BACKGROUND

This paper contributes to the ongoing socio-pragmatic shift in anglicism research by tapping into the current attention for the acquisition of variation in sociolinguistics, sometimes referred to as "developmental sociolinguistics". Particularly, results are presented from an investigation into the development of English lexical resources in the speech of Belgian Dutch preadolescents. In **Section 2**, we discuss the creation of our preadolescent corpus, elaborating on the community of practice under scrutiny and the chosen fieldwork method of sociolinguistic interviews. Next, we introduce our protocol to identify and classify English insertions in the corpus. **Section 3** presents the results of this paper, the implications of which will be discussed in **Section 4**. First, we provide the necessary background on anglicism research and developmental sociolinguistics (1.1), followed by an outline of the Dutch-English contact situation in Flanders (1.2).

1.1 Anglicism Research Featuring Developmental Sociolinguistics
In the current climate of ever-increasing globalization (Blommaert, 2010), English continues to diffuse into Europe's linguistic landscape, stimulating a tradition of anglicism research that focuses

on the resulting English borrowings in European domestic languages. Initially, a large part of the studies in the field have charted all possible manifestations or "types" of English and proposed classification strategies and algorithms to answer the question of how English should be defined and counted in receptor language corpora. For instance, Onysko and Winter-Froemel (2011) introduced the labels of "catachrestic" (~necessary or unavoidable) and "non-catachrestic" (~non necessary, avoidable) loanwords to account for the absence (~catachrestic) or presence (~non-catachrestic) of receptor language alternatives. Other strategies to classify anglicisms involve etymology, adaptation to receptor language morphology or phonology and listedness in dictionaries (Gerritsen et al., 2007; Roberts et al., 2021). Moreover, English loans are usually tagged for part of speech and for the semantic field they belong to (e.g., Yang, 1990). As such, English insertions are shown to be clustered around i.a. *IT/gaming*, *business* and *sports*, thus being labeled as "English-prone" semantic fields (Onysko, 2007; Balteiro, 2018; Hunt, 2019).

In a second wave of anglicism research, the focal point was shifted to the social and pragmatic meanings assigned to these manifestations of English. Vaattovaara and Peterson (2019), for instance, rely on a mixed methods approach to uncover the indexical link between English swearing in Finnish and "urbanity". A more pragmatic perspective was taken in Andersen (2017) who observed a change in discourse function for the Norwegian anglicism *jobb* from negative to more neutral or positive. This "socio-pragmatic turn" in anglicism research (cf. Andersen et al., 2017) foregrounds the perspective of the language user in a quest for the how and why of the language user's choice (not) to include English foreign material in specific contexts, thus achieving a certain social or pragmatic outcome (Onysko and Winter-Froemel, 2011; Peterson and Beers Fägersten, 2018).

For one thing, studies adopting this socio-pragmatic focus reveal how English is a clear youth language marker occurring in game talk, social media interaction and rap- and hip-hop lyrics, making the teenager one of the prototypical English language users (Leppänen, 2007; Pennycook, 2007). However, it is still unclear how and when these teenagers get to the point of using English in the first place: what happens in preadolescence, viz. the transition from childhood to adolescence, remains undocumented. This issue provides the point of departure for this paper and will be tackled against the background of the emerging field of "developmental sociolinguistics".

Developmental sociolinguistics is an upcoming interdisciplinary framework situated at the intersection of language acquisition and sociolinguistics. The term was first used by Entwisle (1966) when investigating how children learn to identify and employ socially meaningful variation patterns (De Vogelaer et al., 2017). The vast majority of studies in the field provide insight into the acquisition of social meaning, studying the stratification of standard and vernacular use according to children's socio-demographic profiles. As concerns age, results point to the standard being used more by younger children (Roberts, 1994) with an increase of non-standard variants when growing older (Smith et al., 2007). Next, the findings on gender patterns are largely inconsistent (Smith and Durham,

2019) and can be categorized into three groups, according to their contradictory conclusions (Nardy et al., 2013): (1) girls using more standard variants than boys (Romaine 1984), (2) boys oppositely using more standard variants than girls (Chevrot, 1991; Roberts, 1997); and (3) absence of a gender effect (Chabanal, 2001; Foulkes et al., 2001). Finally, concerning socio-demographic background, scholars observe a tendency of higher standard use for higher social class (Chevrot et al., 2000; Nardy, 2008).

In this paper, we apply the methods and principles of developmental sociolinguistics, usually targeting standard and vernacular variation, to the unresolved questions about English use by (preadolescent) children. An optimal setting for our intents can be found in Flanders, where contact between English and Dutch is fundamentally present and, moreover, well documented.

1.2 Contact Between English and Belgian Dutch

(Belgian) Dutch is one of Belgium's official national languages, alongside French and German. It is primarily spoken in Flanders, the northern and most populated part of the country. Where Brussels, the capital, has an intensive business, political and also personal-based contact situation between (native) speakers of English, French and Dutch (Mettewie and Janssens, 2006), Flanders has no such level of bi- or multilingual communication: despite expected domain loss in typical areas such as international business and tertiary education, contact with English remains primarily remote and indirect (Rys et al., 2019). It is notably through mass media such as the Internet, pop music and English spoken TV shows and films that contact with English is established (Booij, 2001). Given this indirect nature of the contact situation (see Onysko, 2009), the influence of English on Belgian Dutch is largely limited to the introduction of loanwords and -phrases to the receptor language lexicon.

Recent studies taking a production perspective and targeting Belgian Dutch adults corroborate these findings. A case in point is Zenner and Van De Mieroop (2017) who analyzed the language use of three participants in a Dutch reality TV show. Their use of English is characterized as having a locally emergent and highly dynamic social meaning, amongst others indexing masculinity and brotherhood. Alongside face-to-face conversations, computer-mediated communication too has been the subject of inquiry, with growing attention for the analysis of tweets. For instance, research on the *ooit/ever* construction (Zenner et al., 2018) and the deconstructionalization of the *pimped ride* (De Pascale et al., 2022 forthc., this issue) has demonstrated a developing creativity with English on Dutch Twitter.

Turning to English as a youth language marker, primarily the study of De Decker and Vandekerckhove (2012) is worth mentioning. In their analysis of more than 200,000 MSN chat messages, they identified at least one English insertion in 13.3% of the posts. Interestingly, markedly different results are found in Zenner and Van De Mieroop (2021)'s study on parent-child interactions involving dinner table conversations between 16 parents and 18 children aged 1 to 7, complemented by

sociolinguistic interviews with the parents. Here, less than 1% of the utterances contain minimally one English word and parents, furthermore, report to have no socialization aim towards English. Between preschoolers' non-production (in parent-child interactions) and teenagers' considerable production of English, a transition must take place.

What is more, the results of SLA research in Flanders provide further evidence for the presence of this as of yet undocumented transition. Although English tuition only starts at the second year of secondary school (at age 13), Flemish children already have high receptive English vocabulary knowledge before the age of 12, viz. prior to formal instruction[1] (Peters et al., 2019; Puimège and Peters 2019; De Wilde et al., 2020; De Wilde et al., 2021). The respondents' level of English proficiency is, however, largely idiosyncratic, depending on contextual factors and being related to the types of English input teenagers have access to (ibid.). Indeed, Bollansée et al. (2021) found a positive correlation between productive word knowledge of English and the frequency of playing video games. Since boys are more regularly engaged in English-themed gaming activities than girls (e.g., Kuppens, 2010), this gaming pattern additionally points to a gender effect in overall English use, with boys having a higher expected English production than girls. Following the aforementioned studies, it is then likely that since children younger than 12, through high extramural exposure, receptively understand a lot of English words, they would also produce them. Whether this assumption holds true is what this study aims to address.

1.3 Research Questions

This paper makes a case for applying a developmental sociolinguistic perspective to socio-pragmatic studies in anglicism research, aiming to uncover when, why and how preadolescents use English lexical resources in Belgian Dutch. Three research questions are addressed:

RQ1. How many English insertions do we find in the language use of Belgian Dutch preadolescents overall, taking into account the type of English used?

Following the aforementioned work of Zenner and Van De Mieroop (2021) involving preschoolers in parent-child interactions (less than 1% of English) and De Decker and Vandekerckhove (2012) targeting adolescents (13.3% of English), we hypothesize to find an intermediary frequency, situated in between those two percentages, for the preadolescent age group. As previous studies have shown varying results for different types of English, we insist on categorizing the English elements found in terms of e.g. "avoidability" (cf. catachrestic and non-catachrestic loans, Onysko and Winter-Froemel, 2011) or listedness in dictionaries (Roberts et al., 2021).

RQ2. To what extent do we find stratification by age and gender in Belgian Dutch preadolescents' use of English lexical resources in Dutch, taking into account the type of English used?

For the age trajectory, we advance two conflicting hypotheses: either preadolescents evolve in using English gradually with age, as was found for non-standard variants in previous developmental sociolinguistic studies (Roberts, 1994; Smith et al., 2007); or, following the idiosyncratic nature of English vocabulary learning (i.a. Puimège and Peters, 2019; De Wilde et al., 2021), the preadolescent age group presents a high amount of individual variation with no clear age pattern to be identified. Similarly, the impact of gender on the use of English insertions in the studied transition period is difficult to assess, given the conflicting results in earlier work in developmental sociolinguistics (see **Section 1.1**). We could, nonetheless, cautiously hypothesize to find more English lexical material in boys' speech because of their prototypical gamer's image and the earlier demonstrated importance of gaming (i.a. Bollansée et al., 2021, **Section 1.2**).

RQ3. How can an in-depth analysis of English lexemes in well-targeted individual users help explain the patterns found in RQ1 and RQ2?

A more fine-grained analysis of individual users can possibly throw light on additional parameters, aside from gender and age, steering the insertion of English lexical material in Belgian Dutch. Taking into account the English-prone semantic fields discussed in **Section 1.1**, the amount of English lexical material can for instance also depend on the topic that is discussed. We therefore expect to see a topic effect and resulting English hotspots for e.g. English-prone *gaming* and *sports*.

2 METHODOLOGY

The research questions are addressed through sociolinguistic fieldwork in a cohesive community of practice (Lave and Wenger, 1991, see **Section 2.1**) where we conducted sociolinguistic interviews (**Section 2.2**). The resulting corpus of 26 hours and 28,998 utterances, 15,465 of which originating from our preadolescent respondents, was transcribed (**Section 2.3**) and then mined for English insertions for which we created a phased identification protocol (**Section 2.4**).

2.1 Community of Practice

Respondents were recruited in a hockey club in a middle-sized town in Flanders, Belgium. We included children from all boys' and girls' teams with ages ranging between 6 and 13 years old (M age = 9;8[2], SD = 1;8). The resulting sample consists of 26

---

[1]For secondary school students in other countries, studies have shown similar high English proficiency rates, yet equally reveal L1 transfer effects (see Lorenz et al., 2021).

[2]In this paper, age is displayed in "years(;)months", following the CHAT conventions of the CHILDES project (MacWhinney, 2000, see **Section 2.3**).

preadolescent respondents of whom 12 are boys (M age = 9;2, SD = 1;7) and 14 are girls (M age = 10;2, SD = 1;8; see **Table 1**, **Section 2.3**). The respondents are mainly monolingual speakers of Dutch[3] with an educational track situated between the first year of primary school and the first year of secondary school. As a result, with two exceptions, these preadolescents have not had any formal instruction of English yet[4] (cf. **Section 1.2**). All children provided assent to participate in the study and were given partial disclosure of the research purpose, parents provided informed consent and were given full disclosure[5].

The reason behind choosing a hockey club to recruit respondents is threefold: firstly, a sports club in general, as opposed to a school, is a more informal and therefore more favorable context to collect production data for a youth language phenomenon. Secondly, hockey players in Flanders traditionally have middle-to upper-class backgrounds which allows us to keep the SES variable in this study relatively stable. Thirdly, the hockey club presents an ideal community of practice that is both cohesive, because children all join hockey practice, and dynamic, since social networks do not completely overlap given that the children attend different schools.

The above-described community of practice was recruited for a larger research project investigating the development of social meaning of English. A total of 114 hours of conversational data was collected in the project, both in individual sessions as in group interactions, amounting to 7 hours of data per respondent. In this paper, we zoom in on a subset of the data collection that consists of individual sociolinguistic interviews.

2.2 Sociolinguistic Interviews

Our goal was to track English insertions through preadolescents' unmonitored, casual speech. Since it is the very essence of the sociolinguistic interview to elicit this type of vernacular-like speech, we decided to rely on this staple of Labovian sociolinguistic fieldwork methodology. Prototypically, the sociolinguistic interview involves a near-natural one-on-one casual conversation on everyday topics and emotional memories between a researcher and a language user from a local community (Meyerhoff, 2016). In the strict Labovian tradition (Labov, 1984), the sociolinguistic interview consists of multiple components including the actual interview questions, a reading task, a word list and a list of minimal pairs, each targeting the same vernacular variable (Becker,

2013). In the remainder of this paper, we use "sociolinguistic interview" to refer to the looser definition of making one-on-one recordings of only the interview part of the Labovian format.

As for the practical details of the study, respondents individually participated in the sociolinguistic interviews in March and April 2021. The interviews were conducted online[6] through MS Teams, on a laptop or computer that was set up by a parent in a separate room. All preadolescent children were interviewed by the same 22-year-old Belgian Dutch-speaking female researcher.

For the content and structure of the sociolinguistic interview, we started from the traditional Labovian version (1984) which was then adapted in two ways. On the one hand, we made changes in light of our research goal (identifying English). We designed a semi-structured interview with topic control including i.a. English-prone *IT/gaming* and *sports* and more Dutch-prone *classroom stories* and *leisure activities*. The choice for these semantic fields was made by combining information from an extensive literature review on English loanwords (cf. Onysko, 2007; Balteiro, 2018; Hunt, 2019) and through a large-scale pretest survey targeting perceptions of and attitudes towards English insertions in Belgian Dutch (see Schuring et al., 2021). On the other hand, we tailored the research method to our "young" target group by (1) including personal narratives of the researcher, as is proposed in the Conversational Map Eliciting Procedure (Peterson and McCabe, 1983); and by (2) reworking and updating traditional sociolinguistic questions (cf. Holmes-Elliott, 2021). As such, we transposed Labov's near-death experience narrative (1972, 1984) into a more child-friendly gaming context: "have you ever been in a very dangerous situation when you were gaming?" These procedures resulted in a protocol in which the questions gradually become more personal and challenging, starting with social and demographic information, moving on to emotional experiences and finishing with hypotheticals.

2.3 Corpus and Annotation

The resulting corpus consists of 26 one-hour sociolinguistic interviews containing a total of 28,998 utterances. Half of the utterances were produced by the researcher and will be disregarded for further analysis[7]. The remaining 15,465 preadolescent utterances are the core research object for this study. **Table 1** provides an overview of the corpus in terms of Child ID and alias, Age (M = 9;8, SD = 1;8), Number of Utterances (M = 595, SD = 113) and Mean Utterance Length (in words per utterance, M = 8.92, SD = 1.38).

The corpus was manually transcribed and annotated following the CHAT conventions of the CHILDES project (MacWhinney,

---

[3]Two participants, brother and sister, have been raised bilingually in Dutch and French. Where necessary, this is taken into account in the analysis of the data.

[4]Except for the two oldest girls in our sample (Girl[13] and Girl[14], cf. **Table 1**) whose schools organize English courses as of the first year of secondary school. We factored this into the analysis of the data.

[5]More specifically, respondents, after being asked for "assent", were told they were participating in "linguistic research" leaving our goal to monitor English insertions unclear. Parents had full disclosure from the start and provided informed consent, but were repeatedly asked not to share the research purpose with their children. All personal information in this paper has been pseudonymized following the ethics application for this study and the larger ongoing research project, approved by the Social and Societal Ethics Committee "SMEC" at KU Leuven, approval number G-2020-1998-R5.

[6]After extensive pretesting of both the online and the traditional face-to-face format, we decided to work with online data collection. Aside from the fact that the online format was more practical to organize during the pandemic, a recorded meeting was considered less intrusive than a face-to-face interview. The main reason for this is the discreteness of the camera in the online setting, resulting in a reduced "Observer's paradox" (Labov, 1972).

[7]Needless to say, priming effects have been checked for the entire corpus.

Child ID	Alias	Age	Number of utterances	Mean utterance length
Boy[1]	Nathan	7;5	518	8.23
Boy[2]	Thomas	7;7	473	6.48
Boy[3]	Leon	7;7	519	7.25
Boy[4]	Finn	8;2	587	9.58
Boy[5]	Adam	8;4	597	7.22
Boy[6]	David	8;7	509	6.75
Boy[7]	Kobe	8;11	533	9.23
Boy[8]	Simon	9;3	529	7.60
Boy[9]	Elias	10;10	693	10.39
Boy[10]	Noah	11;5	506	9.02
Boy[11]	Max	11;6	647	9.34
Boy[12]	Victor	12;1	500	7.92
Girl[1]	Alice	6;6	298	9.37
Girl[2]	Zoë	8;11	667	7.99
Girl[3]	Jade	8;11	734	8.48
Girl[4]	Rosalie	8;11	733	8.25
Girl[5]	Laura	9;1	521	6.84
Girl[6]	Stella	9;9	579	7.69
Girl[7]	Lily	9;10	637	10.68
Girl[8]	Charlotte	9;10	687	10.37
Girl[9]	June	10;8	547	9.44
Girl[10]	Yasmine	11;4	763	10.50
Girl[11]	Floor	11;6	687	8.60
Girl[12]	Camille	12;5	700	11.03
Girl[13]	Sarah	12;8	801	10.13
Girl[14]	Olivia	13;1	499	10.93
	M	9;8	595	8.92
	SD	1;8	113	1.38

2000). Excerpt (1) presents an example of what the final transcriptions look like:

(1) *INT: &-eumh en welke muziek luister jij graag?
   %eng: *&-umh and what music do you like to listen to?*
   *YAS: ik vin(d) Lizzo wel heel tof.
   %eng: *I really like Lizzo.*
   *YAS: Billie Eilish minder want (.) ik vind da(t) zo droevig ofzo.
   %eng: *Billie Eilish not so much because (.) I find that like so sad.*
   *YAS: ik word daar nie(t) **happy** van.
   %eng: *it doesn't make me **happy**.*

The transcription of an utterance always starts with a three-letter speaker ID preceded by an asterisk (*INT for interviewer, *YAS for Yasmine) or with "%eng", which is used for the English translation of the utterance. In the utterance itself, (…)[8] and &-euh are used for silent and filled pauses respectively and brackets indicate that (parts of) words are not pronounced as a result of phoneme deletion for round brackets and assimilation for square brackets.

[8]The number of points in between the brackets indicates the duration of the pause with (.) corresponding to a short pause and (…) corresponding to a very long pause.

2.4 Identifying English

The identification of English lexical material has been continuously raised as a methodological concern in loanword studies focusing on anglicisms. As mentioned in **Section 1.1**, a variety of algorithms, strategies and parameters have been created to answer the question of what should count as English. Inspired by Roberts et al. (2021), we synthesize the methods proposed in the literature into a phased identification protocol involving three levels (see **Figure 1**): etymology, recognizability and relative markedness (Levinson, 2000; Winter-Froemel, 2013). A decision task on each of the levels results in a classification into four borrowing types, ordered on a scale from low Englishness to high Englishness: no English, unrecognizable English (UE), recognizable unavoidable English (RUE) and recognizable avoidable English (RAE).

As a first step, we need to decide what lexical material from the transcripts is analyzed through the protocol. The point of departure are words or phrases containing at least one free morpheme found in English dictionaries. Hereby, we do not take into account the traditional borrowing-codeswitching dichotomy but instead consider them to be "two outer poles on a continuum" (Zenner and Van De Mieroop, 2021: 8; also see; Backus, 2014). Both English-listed words (*cool*) as phrases (*oh my god*) are therefore referred to as "insertions". Additionally, the fact that only one free morpheme has to be found in the dictionary is a way to account for the possible adaptation to Dutch morphology, especially in the case of verbal inflection for which morphological integration seems to be inevitable (Winter-Froemel, 2008), and for which the result (Dutch *gebottleflipt* vs. *bottle flipped*) would not be listed in English dictionaries.

For each of these units identified as candidates for the identification protocol, we then make a first major division on the etymological level: we verify whether an insertion has English origin according to *het Etymologisch woordenboek van het Nederlands* (Philippa et al., 2018). A "NO" answer on the decision task at this level results in a classification into "No English". As etymological information can be ambiguous, we choose for an inclusive approach by opting for "YES" when at least one of the suggested etymologies is English.

If an insertion proves to be of English origin, it trickles down in the identification protocol to the recognizability level: an English insertion should be recognizable as English to native speakers of Dutch, and this because "the non-Dutch character of a word can only exert influence on the language user's behavior when the expression at issue is identifiable as a non-Dutch word" (Geeraerts and Grondelaers, 2000: 56). Ideally, the non-Dutch character would be operationalized by our respondent's own perception, or by extension by Dutch native speakers' perception, of what characterizes a non-Dutch insertion. However, this would require a series of perception studies that are beyond the scope of this paper. We hence decided to include two parameters that serve as a proxy for the English character of an insertion. The first parameter concerns grapheme-phoneme mapping (Onysko, 2007) and probes whether the spelling of an insertion leads to a transparent and correct pronunciation in the receptor language, following the grapheme-phoneme mapping rules

FIGURE 1 | 3-step English identification protocol.
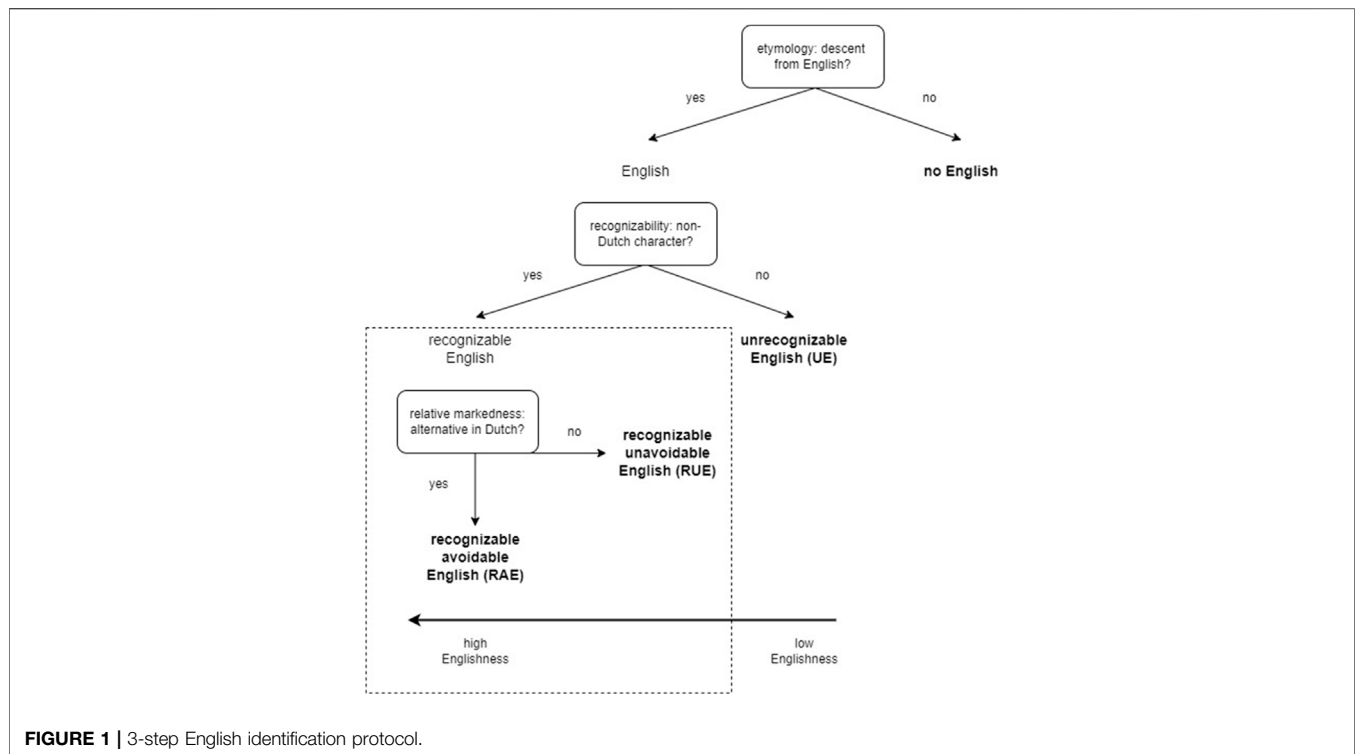
of that same receptor language. For example, an insertion as *challenges* (Excerpt 2) would get a "YES" answer on the recognizability level because Dutch pronunciation (/ˈχaˈlɛnɣəs/) would sound very different from English pronunciation (/ˈtʃæləndʒəz/), whereas the insertion *sport* (/ˈsport/), see Excerpt (3), gets a "NO" answer since naïve Dutch pronunciation matches English pronunciation. The second parameter that leads to English recognizability and proxies the non-Dutch character is the presence of English chargrams (Andersen, 2005; Andersen, 2012; Zenner and Van De Mieroop, 2021). English chargrams can be defined as "a string of n characters within a certain word that have English etymology" such as word-initial *c* in *cornflakes* (Excerpt 4). If both grapheme-phoneme mapping and English chargrams lead to a "NO" answer on this level, the insertion is categorized as unrecognizable English (UE).

(2) *ADA: &-eumh één doet allemaal **challenges** en de andere maakt slijm.
   %eng: *&-umh one does all of these **challenges** and the other makes slime.*
(3) *ELI: &-euh ik kijk nie(t) zo speciaal naar de **sport** op **tv**.
   %eng: *&-umh I don't particularly watch **sports** on television.*
(4) *CHA: je kon **cornflakes** eten croissants enzo.
   %eng: *you could eat **cornflakes** croissants and so on.*
(5) *FLO: ma(ar) buiten **hockey** onthou(d) ik dan ook nie(t) echt veel van mijn **weekends**.
   %eng: *but aside from **hockey** I don't remember much of my **weekends**.*

A recognizable English insertion must finally go through the last step of our protocol which is based on the principle of relative markedness (Levinson, 2000; Winter-Froemel, 2013). This principle corresponds to the labels of "catachrestic (~necessary) and "non-catachrestic"(~non necessary) loanwords (Onysko and Winter-Froemel, 2011, cf. **Section 1.1**) as it accounts for the onomasiological difference between a "marked" insertion because of presence of a Dutch alternative[9] and an "unmarked" insertion because absence of the same. In our protocol, the existence of Dutch alternatives was verified through three Dutch dictionaries (*Van Dale, Het Algemeen Nederlands Woordenboek* (ANW) and *woorden.org*). An insertion was labeled "marked" when at least one of the dictionaries provided a Dutch lexical alternative and, to check for a minimal level of entrenchment (cf. Zenner et al., 2014), when that alternative lexicalization had more than 2,000 hits on Google (Dutch language settings). As a result, words like *hockey* and *weekends* (Excerpt 5, no Dutch alternatives) are categorized as non-marked and get labeled recognizable unavoidable English (RUE), whereas words like *happy* (Excerpt 1, Dutch alternative: *blij*), *challenges* (Excerpt 2, Dutch alternative: *uitdagingen*) and *cornflakes* (Excerpt 4, Dutch alternative: *ontbijtgranen*) get a "YES" answer for relative markedness which gives them the

---

[9]We use "Dutch alternative" to refer to all lexicalizations behaving as near-synonyms, thus following Edmonds and Hirst (2002) who state that the existence of true synonymy is arguable. This implies that the Dutch alternatives inevitably differ from the English borrowings in at least some ways, "varying in their shades of denotation, connotation, implicature, emphasis, or register" (ibid: 107) and also, as is the case for most loanwords, in their degree of specificity (see Backus, 1996).

label of recognizable avoidable English (RAE). Note that in the case of "proper nouns", which we define as a string of words that designates a person, place or thing that has been claimed in the physical world, the decision of relative markedness always leads to a "NO" answer as proper names are *ipso facto* unavoidable insertions.

The above-described identification protocol is applied to all preadolescent utterances in our corpus. In the analysis of the results, we do not take into account unrecognizable English (UE), precisely because of its unrecognizable nature and resulting low Englishness. Instead, in the rest of this paper, we focus on the RUE and RAE borrowing types. All analyses are conducted in R, with the utterance level as a point of departure: for each utterance, we indicate whether or not it includes any instances of RAE or RUE.

3 RESULTS

The results are divided into three sections, corresponding to our three research questions. We first present the overall frequency of English lexical resources in the preadolescent respondents' speech (**Section 3.1**, RQ1), followed by an analysis in terms of stratification by age and gender (**Section 3.2**, RQ2). Finally, we explain the patterns found in **Sections 3.1–3.2** by zooming in on three well-targeted individual language users (**Section 3.3**, RQ3): Kobe (8;11), Max (11;6) and Camille (12;5).

3.1 Overall Frequency of English Lexical Resources

Table 2 presents the overall use of English insertions in our corpus: 9.7% of the utterances contain recognizable English insertions (RAE + RUE). This percentage corresponds to 1,502 utterances out of the corpus total of 15,465. Recognizable unavoidable English (RUE) and recognizable avoidable English (RAE) have a two to one ratio, RUE (6.7%) being significantly more frequent than RAE (3.0%) (results for Wilcoxon rank sum exact test: $W = 28$, $p < 0.001$), for which a large effect (WilcoxonR = 0.7) was found.

The corpus contains 1,695 English tokens and 581 English types, resulting in a type/token ratio of 34.3% (same type/token ratio for RAE and RUE individually). The three most frequent types for RAE are *match* (N=58), *team* (N=23) and *stick* (N=19); the most frequent RUE types are *hockey* (N=127), *computer* (N=48) and *Fortnite* (N=24); for further reflection on these individual types see **Section 4**. It immediately stands out that four of these six frequent insertions are related to sports. This naturally follows from our choice for hockey players as a respondent group and confirms the previous studies' labeling of *sports* as an English-prone semantic field (see **Section 1.1**). A final comment on these numbers relates to the proportion of names in our corpus, such as *Fortnite*[10] (N=24). As mentioned

in **Section 2.4**, all English-inspired proper names figure in the RUE borrowing type. Overall, 216 types or 54.7% of RUE are proper names. As the example of *Fortnite* illustrates, the proper names in the corpus mainly refer to objects and concepts such as *PlayStation* and *For Girls Only*. We chose to retain these names because, although unavoidable, they are still recognizable as being English. In the next section, we investigate how the English insertions discussed above are distributed across respondents.

3.2 Stratification

Figure 2 shows the relative utterance-based frequency of English insertions per type (RUE and RAE) and per respondent. A table version of the plot can be found in **Supplementary Material S1**. The y-axis represents the percentage of utterances in the corpus containing (recognizable) English insertions and ranges from 0 to 15%. The x-axis portrays the individual respondents with a code consisting of their alias and corresponding age (in years and months). The youngest respondent is placed on the left of the x-axis, followed by the other respondents in the order of ascending age. Further, the color scheme (both in the graphs as in the x-axis labels) represents the gender parameter[11] with light blue coding for boys (B) and dark blue coding for girls (G). Last, the full line represents the relative frequency of RUE and the dashed line portrays the relative frequency of RAE.

Figure 2 reflects four stratification patterns, relating to (1) the RAE/RUE ratio across respondents, (2) age, (3) gender and (4) outliers. The corresponding descriptive results can be found in **Table 3** below. We refer to **Supplementary Material S2** for an overview of the boxplots.

First, the graph resonates the findings of the difference between RUE (M = 6.7%, SD = 1.8%) and RAE (M = 3.0%, SD = 2.1%), reported in **Section 3.1**. The relative frequency of RUE is higher than the relative frequency of RAE for all respondents (except for Kobe, cf. **Section 3.3.1**). Moreover, the two graphs mostly follow a parallel trajectory, with respondents who generally use proportionally more RUE, also use more RAE and vice versa.

Second, **Figure 2** shows no clear stratification by respondent age, as going from the left on the x-axis (youngest respondents) to the right (oldest respondents), the graphs are roughly stable and show no clear upward trend. A Kruskal Wallis rank sum test based on a categorization into three age groups (6–8 years/o, 9–10 years/o and 11–13 years/o), indeed, shows no stratification by age for RAE ($H(2) = 0.483$, $p > 0.05$), RUE ($H(2) = 0.087$, $p > 0.05$) nor for RAE + RUE ($H(2) = 0.039$, $p > 0.05$).

Third, we discuss gender stratification patterns. Gender does not play a role in the overall English use of our respondents (RUE + RAE, $H(1) = 0.024$, $p > 0.05$). Zooming in on the specific borrowing types, gender shows no significance for RUE ($H(1) = $

---

[10]"Fortnite" is a free and popular online video game with the purpose of being the last survivor on an island.

[11]We followed the gender distinction as it is applied in the community of practice under scrutiny. The hockey club works with the traditional and mandatory gender division in boys' and girls' (competition) teams.

TABLE 2 | Overall use of English insertions—total number of utterances 15,465.

	Utterances containing English insertions		English insertions	
	N	%	tokens	types
RAE	463	3.0	499	186
RUE	1,039	6.7	1,196	395
RAE + RUE	1,502	9.7	1,695	581

FIGURE 2 | Relative frequency of English per type per respondent (utterance-based).
Full line: recognizable unavoidable English (RUE)—Dashed line: recognizable avoidable English (RAE).
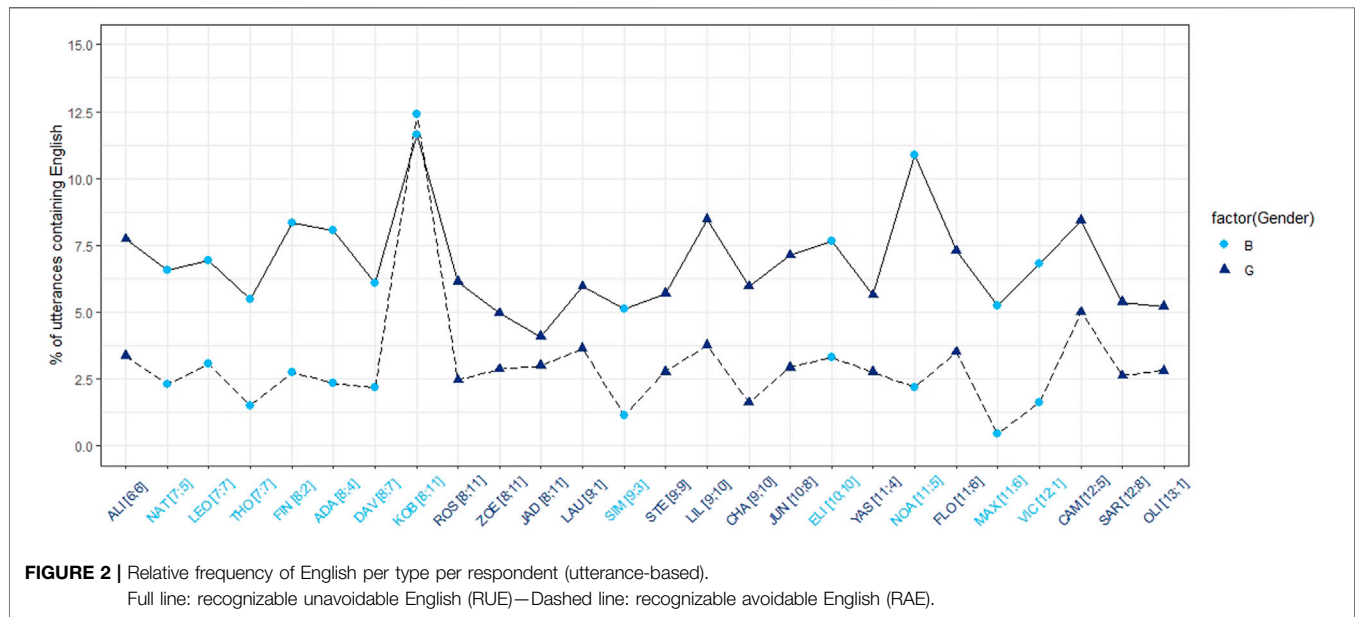
TABLE 3 | Descriptive results for stratification by Age and Gender in % per utterance– *= significant.

Sociodemographic variable	Level	RUE		RAE	
		M	SD	M	SD
Age	6–8	6.91	2.04	3.47	3.00
	9–10	6.57	1.21	2.74	1.01
	11–13	6.86	1.99	2.61	1.33
Gender	M	7.40	2.08	2.93*	3.08
	F	6.29	1.32	3.07*	0.78

1.400, $p > 0.05$) but becomes a factor in explaining the RAE variation ($H(1) = 5.357$, $p = 0.02$, effect size $\eta^2 = 0.182$), with girls (M = 3.1%, SD = 0.8%) using significantly more recognizable avoidable English (RAE) than boys (M = 2.9%, SD = 3.1%). The difference in the means between girls and boys is, however, only 0.2%, indicating that we may not want to overstate the importance of this effect, which also loses its significance when applying Bonferroni correction.[12] In

contrast, support for the presence of a gender effect is found when excluding 8-year-old outlier Kobe from the analysis: the gender stratification becomes more apparent as can be seen in the RAE mean for girls (3.1%, SD = 0.8%) and the new REA mean for boys (2.1%, SD = 0.9%; $H(1) = 8.104$, $p = 0.004$[13], effect size $\eta^2 = 0.296$).

Fourth, the reinforcement of the gender pattern upon Kobe's exclusion already signals the presence and importance of outliers in the sample. The most apparent outlier indeed is Kobe (8;11) who shows the highest relative frequency for both RUE and RAE leading to his profile standing out in **Figure 2**. Otherwise, we observe the opposite pattern when analyzing Max's (11;6) numbers: the lowest frequency for RAE and one of the lowest frequencies for RUE. Additionally, we can identify an outlier in the profile of Camille (12;5) who has the highest RAE frequency and second-highest RUE frequency on the girls' side.

We conclude, answering our second research question, that, in our corpus, RUE is significantly more frequent than RAE for all respondents. At the same time, **Figure 2** and **Table 3** show fairly limited stratification according to traditional socio-demographic parameters, with age not showing any significance and gender reaching robustly significant differences for RAE when outlier Kobe is excluded from the analysis. Indeed, **Figure 2** reveals a

---

[12]In this study, we are performing multiple tests on the same respondent variable, which leads to higher probability of observing significant effects by chance. Bonferroni correction accounts for this by adjusting the significance level: the traditional significance level ($p < 0.05$) is divided by the number of tests performed.

[13]Note that the effect here remains significant upon Bonferroni correction.

number of outliers, individuals just like Kobe, that show a pronounced use of RAE and/or RUE and deserve closer scrutiny in the next section (RQ3).

3.3 An In-Depth Analysis of Three Individual Users

3.3.1 Kobe

Despite being one of the younger respondents in the sample, Kobe (8;11) is the top user of English for both RUE (11.6%) and RAE (12.4%) in our corpus. To explain why this is the case, we performed a topic analysis on all recognizable English insertions (N = 150, token count) identified in Kobe's recorded speech. Our method consisted of tagging the conversation topic in relation to which each insertion was uttered. This was done according to a binary distinction between "gaming-related" and "non-gaming-related" utterances in the conversation. Our choice for this distinction was based on gaming being a highly English-prone semantic field, following the previous research discussed in **Section 1.2** (cf. Puimège and Peters, 2019; De Wilde et al., 2020; De Wilde et al., 2021; Bollansée et al., 2021) and following insights from a qualitative exploration of Kobe's interview. The results of the topic analysis, as well as an account for the total number of tokens for RAE and RUE are given in **Table 4**.

Table 4 shows that out of a total of 150 English tokens identified in Kobe's speech, 103 tokens or 68.7% were uttered when talking about gaming. The percentage even goes up if we only consider RAE: 62 out of 81 tokens, corresponding to 76.5%. Excerpt (6) clarifies what kind of gaming-related insertions we are referring to. Kobe is talking about his favorite video game "Fortnite":

(6) *KOB: der is ook zo iemand een **boss** in de **game**.
%eng: *there is also like someone a **boss** in the **game**.*
*KOB: en der zijn twee (.) &-euh aan elke toren is er zo een **boss**.
%eng: *and there are two (.) &-uh at each tower there is like a **boss**.*
*KOB: en ja die kan je dan **killen**.
%eng: *and yes then you can **kill** him.*
*KOB: want der is ook in **Sweety Sand(s)** is er zo iemand da(t) jou altijd iets gratis geeft.
%eng: *because there is also in **Sweety Sands** there is like someone that always gives you something for free.*
*KOB: soms geeft die jou ook een **medic weapon**.
%eng: *sometimes he also gives you a **medic weapon**.*

The excerpt confirms Kobe's frequent use of recognizable avoidable English when he says *boss* (Dutch alternative: *baas*), *game* (Dutch alternative: *spel*), *killen* (Dutch alternatives: *vermoorden/doden*) and *medic weapon* (Dutch alternative: *medisch wapen*). *Sweety Sands* is the name of a "location" in Fortnite and therefore gets the label of RUE. What is more, the excerpts reveals that, in general, Kobe is a frequent gamer which, in turn, is confirmed by Kobe's further answers to the questions in

TABLE 4 | Kobe.

Type of English	Tokens	Gaming related		Not gaming related	
		n	%	n	%
RAE	81	62	76.5	19	23.5
RUE	69	41	59.4	28	40.6
RAE + RUE	150	103	68.7	47	31.3

the sociolinguistic interview. Apart from explicitly mentioning that he likes gaming and that he games for several hours a day, he also steers the conversation towards the gaming topic, even when he is asked a gaming-neutral question, as can be seen in Excerpt (7):

(7) *INT: en maak jij ook wel es ruzie (.) me(t) jouw broer?
%eng: *and do you sometimes argue (.) with your brother?*
*KOB: &-euh ja veel.
%eng: *&-uh yes a lot.*
*INT: ja en waar gaat dat dan over?
*eng: *yes and what is it about then?*
*KOB: omdat die da(t) wapen eerst wilt enzo.
%eng: *because he wants that weapon first and so on.*
*KOB: want ik had ne keer ik was daar alleen geland en ik heb da(t) per ongeluk **gedropt**.
%eng: *because I had once I landed there once alone and I **dropped** it by accident.*
*KOB: en dan had die da(t) die gouden **shot gun** en dan wouk [: wou ik] da(t) terug.
%eng: *and then he had that that golden **shot gun** and then I wanted it back.*

Kobe's tendency to talk about games all the time (namely in 196 out of his 533 utterances), accompanied by his high frequent English use when doing so, implies a strong topic effect of gaming. This could explain why Kobe is an outlier, producing a much high number of English insertions than the other respondents in the sample. To verify if the topic effect of gaming can indeed account for the variation, we turn to Max (11;6).

3.3.2 Max

Max is nearly 3 years older (11;6) than Kobe (8;11), yet uses very few English insertions (both RAE and RUE). We performed the same topic analysis as described in 3.3.1, which led to the following results:

Table 5 shows that only 10 tokens, or 22.2%, of Max's English insertions were produced when discussing games. Comparing this to Kobe's 68.7% and knowing that Max did not talk much about gaming (cf. in only 45 utterances *vs.* Kobe's 196 utterances), we would be inclined to conclude that talking about games is indeed a predictor for the amount of English lexical material in preadolescents' speech. However, this statement has to be nuanced since Max talked about gaming a little bit in the sociolinguistic interview, yet when doing so, did not produce English insertions, as becomes clear in Excerpt (8):

TABLE 5 | Max.

Type of English	Tokens	Gaming related		Not gaming related	
		n	%	n	%
RAE	6	1	16.7	5	83.3
RUE	39	9	23.1	30	76.9
RAE + RUE	45	10	22.2	35	77.8

(8) *MAX: &-euh een pinkend hartje een spelletje zowa(t) met een vogel da(t) een ja (.) er zijn zo allemaal streepjes.
   %eng: *&-uh a pounding heart a game with like with a bird that a yes (.) there are like all of these stripes.*
   *MAX: en dan zijn er een paar streepjes gevuld en ééntje is over.
   %eng: *and then there are a couple of filled stripes and one is left.*
   *MAX: en dan is er nog zo één blokje vrij en die moe(t) je dan na(ar) boven naar onder na(ar) boven naar onder en zo tot je dood bent.
   %eng: *and then there is only one empty block and you have to (get it) up and down up and down and so on until you're dead.*

In sum, when preadolescent children talk about gaming, they do not necessarily use a lot of English. Max talks about gaming but solely uses Dutch in doing so. This is probably due to the fact that Max's parents do not allow him to play games (see Excerpt 9), which does not give him the opportunity to get familiar with the ingroup gaming code.

(9) * INT: **game** jij ook wel es?
   %eng: *do you sometimes play **games**?*
   *MAX: ah nee mijn mama en papa zijn daar tegen.
   %eng: *ah no my mom and dad are against that.*
   *MAX: ma(ar) ik ik zou ik vin(d) het leuker zonder **gamen**.
   %eng: *but I I would I find it more fun without **gaming**.*

Therefore, it would seem that rather than only a topic effect of gaming, English use is also connected to frequently engaging in gaming activities. We would then find high-frequent English use for high-frequent gamers. Camille (12;5) has the ideal profile to check whether this assumption holds true.

3.3.3 Camille

Camille shows the highest RAE frequency and second-highest RUE frequency of the girls, of which she, herself is one of the oldest (12;5). From the sociolinguistic interview, we know she likes to play video games like *Minecraft* and *Mario Kart*. She also does this frequently and plays together with her friends. **Table 6** presents the topic analysis (cf. **Section 3.3.1**) in terms of English tokens produced in gaming- and not-gaming-related utterances.

Contrary to our expectations, merely 34.0% (N = 35) of Camille's insertions were produced when she was talking about video games (97 utterances). This percentage is only half of Kobe's 68.7% (**Table 4**) and is situated closer to Max's

22.2% (**Table 5**). From these numbers, it appears that "frequently playing games" cannot entirely account for Camille's peak in English use. In what follows, we first throw light on Camille's low frequency of gaming-related English insertions after which we provide an alternative hypothesis to account for her outlier profile.

Although Camille frequently plays games, she does not produce a large number of gaming-related English insertions. Camille provides a possible explanation for this herself, as Excerpt (10) illustrates:

(10) *INT: en **game** jij ook wel es?
   %eng: *and do you sometimes **play games**?*
   *CAM: ja ik heb een **Nintendo Switch**.
   %eng: *yes I have a **Nintendo Switch**.*
   *CAM: die heb ik gekregen in quarantaine.
   %eng: *I got it during quarantine.*
   *CAM: &-euh omda(t) je omda(t) ik mij toch redelijk snel verveel enzo.
   %eng: *&-uh because you because I get bored fairly quickly and all.*
   *CAM: daar heb ik wel een aantal spelletjes op staan.
   %eng: *I do have some games on there.*
   *CAM: ik speel vaak me(t) mijn vrienden **Mario Kart**.
   %eng: *I often play **Mario Kart** with my friends.*
   *CAM: en ja (.) tis nie(t) da(t) ik zo echt serieus **game** ofzo.
   %eng: *and yes (.) it's not like I'm really serious about **gaming** or anything.*
   *CAM: ma(ar) ik vind da(t) wel leuk gewoon als ontspanning.
   %eng: *but I like that just to relax.*

Since Camille states she's not too "serious about gaming", her not being committed to gaming in the way for example Kobe is, can presumably account for her lower gaming-related English use.

This still leaves us with Camille's 103 English tokens to discuss, a number that is already closer to Kobe's 150 tokens, yet not to the same extent connected to gaming. A clarification for this can be found in the nature of the English insertions Camille produced. Consider Excerpts (11) and (12) on this account:

(11) *INT: en heb jet [: je het] ondertussen wel al uitgepraat?
   %eng: *and have you talked things out by now?*
   *CAM: ja ma(ar) das [: dat is] zo die heeft da(t) al zo vaker gedaan he.
   %eng: *yes but that's like she has done that before hey.*

TABLE 6 | Camille.

Type of English	Tokens	Gaming related		Not gaming related	
		n	%	n	%
RAE	37	10	27.0	27	73.0
RUE	66	25	37.9	41	62.1
RAE + RUE	103	35	34.0	68	66.0

*CAM: ma(ar) dan me(t) kleinere dingen zo echt zo **last minute** zo nee zeggen ofzo.

%eng: *but then with smaller things like really like **last minute** like say no or something.*

(12) *INT: en ik vroeg mij af of (.) jij ook nog weet wa(t) je dit **weekend** allemaal gedaan hebt?

%eng: *and I was wondering if (.) you also still remember what you did during the **weekend**?*

*CAM: &-euh ja ik ben zeg maar met die vriendin met Fara ben ik naar Antwerpen geweest.

%eng: *&-uh yes I went like with that friend with Fara I went to Antwerp.*

*CAM: dus zaterdag waren wij in Antwerpen.

%eng: *so on Saturday we were in Antwerp.*

*CAM: en dan hebben wij daar zo een beetje gaan **shoppen enzo**.

%eng: *and then we have done like a little bit of **shopping and stuff**.*

In Excerpt (11), Camille is talking about an argument she had with her friend about participating in a summer camp, stating that her friend decided at the *last minute* not to come, which made her quite angry. Next, in Excerpt (12), Camille discusses her weekend in which she did a shopping visit to Antwerp, together with her friend Fara. Both *last minute* and *shopping* were produced when talking about stereotypically girl-oriented[14] activities. In fact, 37.8% (14 out of 37) of Camille's RAE tokens, and 9.1% (6 out of 66) of her RUE tokens, can be traced back to these girl-oriented topics, with additional English insertions like *pony* (RUE) and *playbacken*, *tie dyen* and *slash* (RAE). The peak in English use for Camille would therefore not result from a topic effect of gaming alone, nor would it stem solely from her frequently playing video games. Camille's high frequent use of English additionally seems connected to her being a preadolescent girl, talking regularly about girl-oriented topics.

To corroborate this, we briefly discuss two other girls in the sample, Charlotte (9;10) and Sarah (12;8), who report on an argument with their friends (Excerpt 13) and on the boys' extracurricular activities on the playground (Excerpt 14):

(13) *CHA: en nu gaan die ruzies zo meestal over van "bemoeit u me(t) uw eigen zaken" ofzo.

%eng: *and now those arguments are like usually about like "mind your own business" or something.*

*CHA: en dan beginnen wij te **diss(en)** te(gen) allez beginnen wij zo scheldwoorden tegen en elkaar te zeggen.

%eng: *and then we start **dissing** to well we start saying like swear words to each other.*

*CHA: omda(t) &-euh ja wij denken dan van "we gaan (i) kga mij we gaan ons nie(t) laten doen" enal.

%eng: *because &-uh yes we then think like "we're going to I'm not letting me we're not letting us be pushed around" and all.*

(14) *SAR: en &-eumh hij had zijn vrienden hadden daar zo op **gebottleflipt**.

%eng: *and &-umh he had his friends had like **bottle flipped** on that.*

*SAR: das [: dat is] zo me(t) waterflesje en dan moet ge zo draaien dat da(t) zo blijft staan ofzo.

%eng: *that's like with a water bottle and then you have to like turn so that it like stays up or something.*

Here, Charlotte and Sarah too talk about girls' activities. In Excerpt (13), Charlotte uses the word *dissen* (*to diss*) to explain how an argument with her friends usually ensues. *Dissen/To diss* has a slang origin and is an informal way of saying you disrespect someone by insulting them in a certain way (cf. Cambridge Dictionary, 2021). The English verb has clearly made its way into Dutch where it has been adapted, presumedly inevitably (cf. Winter-Froemel, 2008), to Dutch verbal inflection. In Excerpt (14), the English verb *to bottle flip* underwent the same morphological adaptation to Dutch. Sarah is talking about the boys, including her brother, and informs the researcher about their activities on the playground: most often they *bottle flip*. Remarkably, this gossipy statement is immediately followed by a detailed explanation of the verb. The same happened in the conversation with Charlotte (Excerpt 13) who even stopped in the middle of the word *dissen* and converted to an explanation "saying swear words to each other". The English insertions in the excerpts thus interestingly get *flagged*[15], i.e. marked as being foreign material in the receptor language (Levendis and Calude, 2019). Setting this aside, Excerpts (13) and (14) seem to confirm an additional topic effect of girl-oriented activities, where high girlishness leads to more English use.

4 DISCUSSION AND CONCLUSION

In this paper we investigated preadolescents' developing use of English lexical resources in Belgian Dutch. In our effort to chart the largely undocumented sociolinguistic transition process from childhood to adolescence, we addressed three research questions: RQ1. How many English insertions do we find, taking into account the type of English used?; RQ2. To what extent is the use of these English insertions stratified by age and gender?; and RQ3. Can the patterns found in RQ1 and RQ2 be explained through an in-depth analysis of Kobe (8;11), Max (11;6) and Camille (12;5)?

---

[14]The division in girl- and boy-oriented activities, we are aware, is in need of more nuance and research and would benefit from being connected to the ongoing societal debate on gender stereotyping.

[15]As Levendis and Calude (2019 : 1) note, flagging can be done in multiple ways including translating/explaining a loan, or using bold face font or italics to demarcate it from surrounding discourse.

The corpus of 15,465 preadolescent utterances was mined for English insertions according to our phased identification protocol, focusing on recognizable unavoidable English (RUE) and recognizable avoidable English (RAE). Results showed an overall utterance-based English frequency of 9.7% (RQ1) This percentage fits in nicely within the expected trajectory, as it is situated in between the previous finding of less than 1% of English insertions in the family home (Zenner and Van De Mieroop, 2021) and corresponding 13.3% in teenagers' social media interactions (De Decker and Vandekerckhove, 2012). As for stratification (RQ2), we found no age effect, some indication of a gender effect for RAE with girls using it significantly more often than boys, particularly when leaving out outlier Kobe; and high levels of idiosyncrasies. These idiosyncrasies reflect the importance of the gaming topic (Kobe) which, at the same time, is nuanced by the possibility to use only Dutch when talking about gaming (Max) and the presumably additional importance of activities stereotypically oriented toward preadolescent girls (Camille, RQ3). This also points to the distinction between English-prone semantic fields and English-prone topics, a tension we have not explored significantly in our contribution. For instance, the English word *bro* can occur in a Dutch conversation on gaming (an English-prone topic), without itself belonging to the semantic field of gaming. In any case, the importance to distinguish between different types of English is confirmed, since we found that RAE (3.0%) occurred much less frequently than RUE (6.7%).

What can these results reveal about the overall development of the position of English in the lexicon of preadolescents? We advance the hypothesis that English is starting to become a youth language identity marker for our respondents (Leppänen, 2007; Leppänen and Nikula, 2007), with an emerging and varying socio-indexical potential, a process in which one preadolescent seems to be faster than the other. Four arguments derived from our study can support this hypothesis.

Firstly, we did not see an age-related development in cohorts but rather saw a fairly stable use of English across preadolescents disrupted by the presence of several outliers. Secondly, the socio-indexical potential of English is observed in the varying results for the use of English when discussing gaming, which seems to be linked to incipient identity construction. Whether or not a preadolescent uses English when talking about gaming, seems in part to depend on the degree to which the preadolescent aims to assume the "gamer identity". Compare Kobe, an ardent gamer and a clear member of the ingroup (high gaming-related English use), to Camille, also a gamer, but not taking it "too seriously", thus not aiming to become an ingroup member (lower gaming-related English use), and to Max, a clear outsider to the gaming community (near-absence of gaming-related English use). Thirdly, a gender effect seems to be at play, with different socio-indexical attributes associated with English in both groups. This is firstly supported by the significance of RAE frequencies in these two groups (when outlier Kobe is excluded), and further in the stereotypically girl-oriented topics addressed and "soon-to-be teenage girl" identity assumed by the girls. Our gender-related conclusions tie in with the mixed gender results in previous (developmental)

sociolinguistic studies (Sections 1.1 and 1.2, cf. Nardy et al., 2013; Zenner and Van De Mieroop, 2017) and indicate that gender may be created very "locally" in discourse. This certainly requires further research. Lastly, the instances of *flagging* we found in Excerpts (13) and (14) can be interpreted as a developing awareness of the ingroup and outgroup marking potential of English insertions. In providing additional explanation (*flagging*) to their use of English terms (*dissen*, *bottle flippen*), the girls indicate that they believe the researcher is likely not familiar with these terms. This further indicates that they might consider the researcher as an outgroup member. These observations of flagging and outgroup marking seem to be a sign of emerging sociolinguistic competence.

In order to solidify our claim about an emerging and variable socio-indexical potential of using English in Dutch, seeping in at various speeds, further research is needed that additionally allows us to overcome some of the shortcomings of this study. First of all, although we factored in the possibility of lexical competition through our distinction between RUE and RAE, follow-up studies might wish to take the onomasiological perspective further (Geeraerts, 2010; Zenner et al., 2014). Particularly our topic analysis would benefit from a comparison of English insertions with the Dutch words or counterparts they are encountered with, for instance benchmarking the use of *game* against its Dutch counterpart (*video*)*spel*. For one thing, this would allow us to arrive at a more local operationalization of relative markedness in our English identification protocol. Our current operationalization is indeed not airtight as it does not account for the individual perspective of the language user. Consider for example *stick* (as in *hockey stick*), the third-most frequent RAE insertion in the corpus (see **Section 3.1**). *Stick* received the RAE label because of its Dutch alternative *stok*, which was frequent enough to be considered in the study (4,000+ hits on Dutch Google). A more fine-grained analysis of the Google hits suggests nonetheless that *stok* was only used by hockey novices or amateurs. Our respondents are frequent hockey players, however, for whom "*stick*" is probably the only possible lexicalization. The RAE status of *stick* and possibly other insertions as well is thus questionable since it does not account for the respondents' individual worlds and the existence of Dutch alternatives therein. This further supports the fact that specificity might be an important motivation for using English insertions (see Backus, 1996 on specificity as driving factor in Turkish-Dutch interactions).

Another prospect for further studies is aiming to find out what our respondents themselves consider to be "English insertions". Given the instances of flagging we observed (cf. explaining an English insertion to the researcher e.g. *dissen, gebottleflipt*), it is likely that respondents are aware of having produced that insertion, moreover, that they have a perception of what characterizes it as being "English". This points towards the importance of including the respondents' perception in the analysis and is what currently lacks in the recognizability level of our identification protocol. That recognizability level is now based on two proxies, namely "grapheme-phoneme mapping" and "non-Dutch chargrams", which can both be criticized for relying too much on spelling conventions our young respondent group has little affinity with.

Besides, following the diverging sociolinguistic competence of our respondents, some of whom are not yet presenting flagging practices, we could expect that at least part of them are not aware of producing English at all and think they are using Dutch. Again, the identification protocol does not yet include this individual metalinguistic feature. We therefore argue for a center role for perception and metalinguistic awareness (cf. Drager and Kirtley, 2016) in future studies, also factoring in the evolution in children's use of English and the input they rely on in their acquisition (see Leona et al., 2021). These concepts, when thoroughly methodologically developed and implemented, can shed more light on (preadolescent) children's developing production of and socio-indexical attributes awarded to English and, in that sense, on their motivations for using English. Likely candidates of such social-attributes for English are 'coolness' (Garley, 2019) or modernity (Piller, 2001) and motivations could include emotion regulation (Eekhof, 2017) or increasing semantic specificity (Backus, 1996).

As a final avenue for future work, we highlight the importance of including a variety of data types. Our one-sided method to conduct sociolinguistic interviews is indeed not without problems since it can be responsible for the above-mentioned outgroup effect: as the participants in this interview are typically only the researcher and the interviewee, the research setting is characterized as "distanced". Apart from the age difference between the participants (approximately 15 years), the interviewee has never met the researcher who does not classify as a peer group member. In sum, it may be that the overall fairly low proportion of utterances containing English (specifically RAE) in our preadolescent corpus is a result of the one-on-one research context, with the children realizing that the researcher is not part of the ingroup and will not understand the ingroup English-inspired code. This conclusion must, however, be tempered by the fact that, in spite of the distanced research setting, and precisely because of careful preparation and child-friendly changes to the format, respondents clearly felt at ease as they spontaneously started to talk about a range of fairly private subjects: getting into trouble, keeping secrets from their parents and having a little sweetheart. In future research, it would nonetheless still be interesting to include data collection in peer group settings (without the researcher being present) in order to counterbalance the possible English inhibitory effect of the one-on-one research setting.

Although the suggested future work is needed to get a full picture, this study already allows us to draw important conclusions about the development of English use in preadolescents, revealing part of the understudied trajectory from childhood to adolescence and arguing for a so-called socio-pragmatic developmental turn in anglicism research.

DATA AVAILABILITY STATEMENT

The data sets presented in this article are not readily available because of ethical restrictions that could comprise the privacy of the research participants.

ETHICS STATEMENT

This study involving human participants was reviewed and approved by the Social and Societal Ethics Committee (SMEC) at KU Leuven, approval number G-2020-1998-R5. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MS performed the data collection, analyzed the data and is first author of the paper. EZ is the overseer of the research project. She provided critical feedback and helped shape the data collection, analysis and manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.788768/full#supplementary-material

REFERENCES

Andersen, G. (2005). "Assessing Algorithms for Automatic Extraction of Anglicisms in Norwegian Texts," in Paper presented at the Corpus Linguistics Conference, Birmingham, July 2005.

Andersen, G. (2017). A Corpus Study of Pragmatic Adaptation: The Case of the Anglicism [Jobb] in Norwegian. J. Pragmatics 113, 127–143. doi:10.1016/j.pragma.2016.12.015

Andersen, G. (2012). "Chapter 5. Semi-automatic Approaches to Anglicism Detection in Norwegian Corpus Data," in The Anglicization of European

Lexis. Editors C. Furiassi, V. Pulcini, and F. Rodriguez-González (Amsterdam, Philadelphia: John Benjamins), 111–130. doi:10.1075/z.174.09

Andersen, G., Furiassi, C., and Mišić Ilić, B. (2017). The Pragmatic Turn in Studies of Linguistic Borrowing. J. Pragmatics 113, 71–76. doi:10.1016/j.pragma.2017.03.010

Backus, A. (2014). "Towards a Usage-Based Account of Language Change: Implications of Contact Linguistics for Linguistic Theory," in Questioning Language Contact: Limits of Contact, Contact at its Limits. Editor R. Nicolaï (Leiden & Boston: Brill), 91–118.

Backus, A. (1996). Two in One. Bilingual Speech of Turkish Immigrants in The Netherlands. Tilburg: Tilburg University Press.

Balteiro, I. (2018). Oh Wait : English Pragmatic Markers in Spanish Football Chatspeak. *J. Pragmatics* 133, 123–133. doi:10.1016/j.pragma.2018.03.011

Becker, K. (2013). "The Sociolinguistic Interview," in *Data Collection in Sociolinguistics: Methods And Applications*. Editors C. Mallinson, B. Childs, and G. Van Herk (New York: Routledge).

Blommaert, J. (2010). *The Sociolinguistics of Globalization*. Cambridge: Cambridge University Press.

Bollansée, L., Puimège, E., and Peters, E. (2021). "Watch Out! behind You Is the Enemy! an Exploratory Study into the Relationship between Extramural English and Productive Vocabulary Knowledge," in *Pop Culture in Language Education*. Editors V. Werner and F. Tegge (London: Routledge), 199–214. Chapt. 12.

Booij, G. (2001). English as the Lingua Franca of Europe: A Dutch Perspective. *Lingua E Stile* 36, 347–357.

Cambridge Dictionary (2021). Cambridge International Dictionary of English. Available at: https://dictionary.cambridge.org/(Accessed August 13–20, 2021).

Chabanal, D. (2001). L'acquisition de phonèmes variables: influences orthographique et sociolectale. *Travaux de Didactique de FLE* 45, 87–96.

Chevrot, J.-P. (1991). La variation phonétique chez des enfants de 6 à 12 ans: aspects développementaux et incidence des apprentissages alphabétiques. PhD dissertation. Grenoble: Université Stendhal.

Chevrot, J.-P., Beaud, L., and Varga, R. (2000). Developmental Data on a French Sociolinguistic Variable: Post-consonantal Word-Final/R/. *Lang. Change* 12, 295–319. doi:10.1017/s095439450012304x

De Decker, B., and Vandekerckhove, R. (2012). English in Flemish Adolescents' Computer-Mediated Discourse: A Corpus-Based Study. *English World-Wide* 33, 321–352. doi:10.1075/eww.33.3.04ded

De Pascale, S., Pijpops, D., Van de Velde, F., and Zenner, E. (2022). (forthc.). "Reassembling the Pimped Ride: A Case of Deconstructionalization in a Borrowed Expression," in *Englishes in a Globalized World: Exploring Contact Effects on Other Languages*. Editors A. Onysko and P. Siemund. accepted for publication in *Frontiers in Communication*.

G. De Vogelaer and M. Katerbow (Editors) (2017). *Acquiring Sociolinguistic Variation* (Amsterdam: John Benjamins).

De Wilde, V., Brysbaert, M., and Eyckmans, J. (2021). Formal versus Informal L2 Learning. *Stud. Second Lang. Acquis*, 1–25. doi:10.1017/s0272263121000097

De Wilde, V., Brysbaert, M., and Eyckmans, J. (2020). Learning English through Out-Of-School Exposure. Which Levels of Language Proficiency Are Attained and Which Types of Input Are Important? *Bilingualism* 23, 171–185. doi:10.1017/s1366728918001062

Drager, K., and Kirtley, J. (2016). "Awareness, Salience, and Stereotypes in Exemplar-Based Models of Speech Production and Perception," in *Awareness and Control in Sociolinguistic Research*. Editor A. M. Babel (Cambridge: CUP), 1–24.

Edmonds, P., and Hirst, G. (2002). Near-synonymy and Lexical Choice. *Comput. Linguistics* 28, 105–144. doi:10.1162/089120102760173625

Eekhof, L. (2017). Codeswitching as a Form of Emotion Regulation. [unpublished master's thesis]. Utrecht: Utrecht University.

Entwisle, D. R. (1966). Developmental Sociolinguistics: A Comparative Study in Four Subcultural Settings. *Sociometry* 29, 67–84. doi:10.2307/2786011

Foulkes, P., Docherty, G., and Watt, D. (2001). The Emergence of Structured Variation. *Univ. Pa. Working Pap. Linguistics* 73, 67–84.

Garley, M. (2019). 'Do They Know the normal Language?': Language Attitudes and Ideologies in German Hip Hop Culture. *Crit. Multilingualism Stud.*, 7, 93–128.

Geeraerts, D. (2010). "Lexical Variation in Space," in *Language in Space: An International Handbook of Linguistic Variation. Vol. 1, Theories and Methods*. Editors P. Auer, J. E. Schmidt, and A. Lameli (Berlin: Mouton de Gruyter), 821–837.

Geeraerts, D., and Grondelaers, S. (2000). Purism and Fashion. French Influence on Belgian and Netherlandic Dutch. *Belgian J. Linguistics* 13, 53–68. doi:10.1075/bjl.13.04gee

Gerritsen, M., Nickerson, C., Van Hooft, A., van Meurs, F., Nederstigt, U., Starren, M., et al. (2007). English in Product Advertisements in Belgium, France, Germany, the Netherlands and Spain. *World Englishes* 26, 291–315. doi:10.1111/j.1467-971x.2007.00510.x

Holmes-Elliott, S. (2021). Calibrate to Innovate: Community Age Vectors and the Real Time Incrementation of Language Change. *Lang. Soc.* 50, 441–474. doi:10.1017/s0047404520000834

Hunt, J. W. (2019). "Anglicisms in German: Tsunami or Trickle," in *Informalization and Hybridization of Speech Practices: Polylingual Meaning-Making across Domains, Genres, and Media*. Editor A. Koll-Stobbe (Bern: Peter Lang,), 25–58.

Instituut voor de Nederlandse taal (n d) *Algemeen Nederlands Woordenboek* Online. Available at: https://anw.ivdnt.org/(Accessed August 13–20, 2021).

Kuppens, A. H. (2010). Incidental Foreign Language Acquisition from media Exposure. *Learn. Media Tech.* 35, 65–85. doi:10.1080/17439880903561876

Labov, W. (1984). "Field Methods of the Project on Linguistic Change and Variation," in *Language in Use: Readings in Sociolinguistics*. Editors J. Baugh and J. Sherzer (Englewood Cliffs: Prentice-Hall), 28–53.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lave, J., and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: CUP.

Leona, N. L., van Koert, M. J. H., van der Molen, M. W., Rispens, J. E., Tijms, J., and Snellings, P. (2021). Explaining Individual Differences in Young English Language Learners' Vocabulary Knowledge: The Role of Extramural English Exposure and Motivation. *System* 96, 102402–102413. doi:10.1016/j.system.2020.102402

Leppänen, S., and Nikula, T. (2007). Diverse Uses of English in Finnish Society: Discourse-Pragmatic Insights into media, Educational and Business Contexts. *Multilingua* 26, 333–380. doi:10.1515/multi.2007.017

Leppänen, S. (2007). Youth Language in media Contexts: Insights into the Functions of English in Finland. *World Englishes* 26, 149–169. doi:10.1111/j.1467-971x.2007.00499.x

Levendis, K., and Calude, A. (2019). Perception and Flagging of Loanwords - A Diachronic Case-Study of Māori Loanwords in New Zealand English. *Ampersand* 6, 100056. doi:10.1016/j.amper.2019.100056

Levinson, S. C. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature*. Cambridge (MA) & London: The MIT Press.

Lorenz, E., Hasai, Y., and Siemund, P. (2021). Multilingual Lexical Transfer Challenges Monolingual Educational Norms: Not Quite!. *Multilingua* 40, 791–813. doi:10.1515/multi-2021-0014

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah (NJ): Lawrence Erlbaum Associates.

Mettewie, L., and Janssens, R. (2006). "Chapter 5. Language Use and Language Attitudes in Brussels," in *Multilingualism in European Bilingual Contexts*. Editors D. Lasagabaster and A. Huguet (Bristol, Blue Ridge Summit: Multilingual Matters), 117–143. doi:10.21832/9781853599316-007

Meyerhoff, M. (2016). Methods, Innovations and Extensions: Reflections on Half a century of Methodology in Social Dialectology. *J. Sociolinguistics* 20, 431–452. doi:10.1111/josl.12195

Nardy, A. (2008). Acquisition des variables sociolinguistiques entre 2 et 6 ans: facteurs sociologiques et influences des interactions au sein du réseau social. PhD dissertation. Grenoble: Université Stendhal.

Nardy, A., Chevrot, J.-P., and Barbu, S. (2013). The Acquisition of Sociolinguistic Variation: Looking Back and Thinking Ahead. *Linguistics* 51, 255–284. doi:10.1515/ling-2013-0011

Nederlands Woordenboek (n d). woorden.org Online. Available at: https://www.woorden.org/(Accessed August 13–20, 2021).

Onysko, A. (2007). *Anglicisms in German. Borrowing, Lexical Productivity and Written Code-Switching*. Berlin/New York: Walter de Gruyter.

Onysko, A. (2009). "Divergence with a Cause? The Systemic Integration of Anglicisms in German as an Indication of the Intensity of Language Contact," in *Englischer Sprachkontakt in den Varietäten des Deutschen/ English in contact with varieties of German*. Editor F. Pfalzgraf (Frankfurt am Main: Peter Lang,), 53–74.

Onysko, A., and Winter-Froemel, E. (2011). Necessary Loans - Luxury Loans? Exploring the Pragmatic Dimension of Borrowing. *J. Pragmatics* 43, 1550–1567. doi:10.1016/j.pragma.2010.12.004

Pennycook, A. (2007). Language, Localization, and the Real: Hip-Hop and the Global Spread of Authenticity. *J. Lang. Identity Educ.* 6, 101–115. doi:10.1080/15348450701341246

Peters, E., Noreillie, A. S., Heylen, K., Bulté, B., and Desmet, P. (2019). The Impact of Instruction and Out-of-School Exposure to Foreign Language Input on Learners' Vocabulary Knowledge in Two Languages. *Lang. Learn.* 69, 747–782. doi:10.1111/lang.12351

Peterson, C., and McCabe, A. (1983). *Developmental Psycholinguistics: Three Ways of Looking at a Child's Narrative*. New York: Plenum.

Peterson, E., and Beers Fägersten, K. (2018). Introduction to the Special Issue: Linguistic and Pragmatic Outcomes of Contact with English. *J. Pragmat.* 133, 105. doi:10.1016/j.pragma.2018.06.005

Philippa, M., Debrabandere, F., Quak, A., Schoonheim, T. H., and van der Sijs, N. (2018). Etymologisch Woordenboek Van Het Nederlands (ENW). Instituut voor Nederlandse Lexicologie. Available at: https://etymologie.nl/(Accessed August 13–20, 2021).

Piller, I. (2001). Identity Constructions in Multilingual Advertising. *Lang. Soc.* 30, 153–186. doi:10.1017/s0047404501002019

Puimège, E., and Peters, E. (2019). Learners' English Vocabulary Knowledge Prior to Formal Instruction: The Role of Learner-Related and Word-Related Variables. *Lang. Learn.* 69, 943–977. doi:10.1111/lang.12364

Roberts, G., Zenner, E., and Rosseel, L. (2021). Exploring English in TV Product Advertising for Dutch-speaking Children. *World Englishes*, 1–18. doi:10.1111/weng.12576

Roberts, J. (1994). Acquisition of Variable Rules: (T-,d) Deletion and (Ing) Production in Preschool Children. Unpublished PHD dissertation. Philadelphia: University of Pennsylvania.

Roberts, J. (1997). Acquisition of Variable Rules: a Study of (-t, D) Deletion in Preschool Children. *J. Child. Lang.* 24, 351–372. doi:10.1017/s0305000997003073

Romaine, S. (1984). *The Language of Children and Adolescents: The Acquisition of Communicative Competence.* Oxford: Blackwell.

Rys, K., De Caluwe, J., Doerga Misier, S., Balesar, U., Rozenblad, M., Heeringa, W., et al. (2019). *Onderzoeksrapport Staat van het Nederlands. Over de taalkeuzes van Surinamers in het dagelijks leven en meer over die van Nederlanders en Vlamingen.* Den Haag: Taalunie/Amsterdam: Meertens Instituut.

Schuring, M., Zenner, E., and Rosseel, L. (2021). "Perceptions of and Attitudes towards Contact-Induced Variation: The Case of Societal Roles Using English in Dutch," Presented at the 5th Variation and Language Processing Conference (VALP5), Copenhagen, August 27, 2021.

Smith, J., and Durham, M. (2019). *Sociolinguistic Variation in Children's Language. Acquiring Community Norms.* Cambridge: CUP.

Smith, J., Durham, M., and Fortune, L. (2007). 'Mam, My Trousers Is Fa'in doon!': Community, Caregiver and Child in the Acquisition of Variation in a Scottish Dialect. *Lang. Variation Change* 19, 63–99. doi:10.1017/S0954394507070044

Vaattovaara, J., and Peterson, E. (2019). Same Old *Paska* or New *Shit*? on the Stylistic Boundaries and Social Meaning Potentials of Swearing Loanwords in Finnish. *Ampersand* 6, 100057. doi:10.1016/j.amper.2019.100057

Van Dale (n d). Van Dale Online Woordenboeken. Available at: https://www.vandale.nl (Accessed August 13–20, 2021).

Winter-Froemel, E. (2008). "Unpleasant, Unnecessary, Unintelligible? Cognitive and Communicative Criteria for Evaluating Borrowings and Alternative Strategies," in *Anglicisms in Europe. Linguistic Diversity in a Global Context.* Editors R. Fischer and H. Pulaczewska (Cambridge: Cambridge Scholars Publishing), 16–41.

Winter-Froemel, E. (2013). "Formal Variance and Semantic Changes in Borrowing: Integrating Semasiology and Onomasiology," in *New Perspectives On Lexical Borrowing: Onomasiological, Methodological And Phraseological Innovations.* Editors E. Zenner and G. Kristiansen (Berlin, Boston: De Gruyter Mouton), 65–100. doi:10.1515/9781614514305.65

Yang, W. (1990). *Anglizismen im Deutschen.* Tübingen. Max Niemeyer Verlag. doi:10.1515/9783111676159

Zenner, E., Heylen, K., and Van de Velde, F. (2018). Most Borrowable Construction Ever! A Large-Scale Approach to Contact-Induced Pragmatic Change. *J. Pragmatics* 133, 134–149. doi:10.1016/j.pragma.2018.05.004

Zenner, E., Speelman, D., and Geeraerts, D. (2014). Core Vocabulary, Borrowability and Entrenchment. *Dia* 31, 74–105. doi:10.1075/dia.31.1.03zen

Zenner, E., and Van De Mieroop, D. (2021). The (Near) Absence of English in Flemish Dinner Table Conversations. *Appl. Linguistics Rev.* 12, 299–330. doi:10.1515/applirev-2019-0029

Zenner, E., and Van De Mieroop, D. (2017). The Social and Pragmatic Function of English in Weak Contact Situations: Ingroup and Outgroup Marking in the Dutch Reality TV Show Expeditie Robinson. *J. Pragmatics* 113, 77–88. doi:10.1016/j.pragma.2017.02.013

Views on "Good English" and "Nordic Exceptionalism" in Finland

Elizabeth Peterson*

Department of Languages, University of Helsinki, Helsinki, Finland

In the Nordic countries, widespread proficiency in English is positioned as a positive and even critical component of overall global competitiveness and competence. Indeed, maps illustrating who speaks the "best" English in Europe show a swath across the Nordic countries, and the number of people in the Nordic countries claiming proficiency in English is only a few percentage points below those in places such as the UK and Ireland. At the same time, the Nordic countries are routinely listed as the "happiest," the most egalitarian, the most classless, least corrupt, and an epicenter for so-called "tender values." In recent years, there has been a spate of publications highlighting how Nordic exceptionalism carries with it some unfortunate downsides, including the possibility for people to ignore or fail to acknowledge issues such as racism, sexism, and other social inequalities because of the affordance: "But our society is equal." There is a parallel in the use of English. The entrenched notion that "everyone is good at English" overlooks that certain segments of the population—such as the elderly, immigrants and rural inhabitants—do not have the same level of access to the symbolic capital represented through facility in English. In this sense, the use of English presents social/class-based barriers that the national languages do not. This article offers a critique of the social realities relating to the use of English in the Nordic Countries within the context of the social welfare system and "Nordic exceptionalism," focusing mostly on Finland. Making use of examples of discourse in newspapers, previous research and language policy documents, the chapter highlights how aspects of the use of English in Finland parallel other potentially hyped yet unequitable social issues.

Keywords: English as a foreign language, Finland, Nordic countries, education, language attitudes, social welfare model

INTRODUCTION

This submission to the special issue on Englishes in a globalized world offers a critical perspective on values and ideologies about English in a specific setting, Finland, while situating aspects about the use of English within the wider social, economic and political landscape. The submission is thus in line with the overall aims of this collection in offering a critical discussion on the spread and function of English as a global language, including the analysis of public opinions and discourse. Key issues brought to the forefront in this discussion include Finland's role as one of five nation

states comprising what is commonly referred to as "the Nordic" countries of Europe[1], discussed further in section Definitions of Nordic Exceptionalism, along with definitions of what is referred to as "Nordic exceptionalism." In recent years, there has been a spate of positive publicity about the Nordic nations and their success in achieving a high level of equality and "happiness." While these notions are widely propped up as desired and even perceived realities, they nonetheless have been questioned in recent accounts, demonstrating that the Nordic states are not without faults, both historically and in the current era. The common adage that "everyone speaks English in Scandinavia[2]" proves problematic in much the same way, and it is this notion that is explored here, using the concept of Nordic exceptionalism as a window to view the phenomenon of English language use.

The article first lays some groundwork by presenting an overview of the English language profile in the Nordic countries, zeroing in on the particular context of Finland—Finland serving as the main focal point of this discussion. The article then goes into detail describing similarities in the governing and social system of the Nordic countries, as these aspects are considered key elements in understanding the role of English in the Nordic setting, as well as arguments that are raised later in the article. As a final piece of background information, the article offers a summary of literature concerning the colonialism and coloniality surrounding the spread of global English as a foreign language. These elements, too, link to later arguments in the article. The article concludes by offering evidence of the extension of "exceptionalism" to the use of English in Finland in three areas: social class, language attitudes, and social exclusion.

ENGLISH LANGUAGE PROFICIENCY IN THE NORDIC COUNTRIES

The widespread adoption of English as a foreign language in the Nordic countries can be seen as an outcome of several concurrent historical occurrences, associated to a large extent with a push toward modernization and the related social reform that emerged in the second half of the 20th century (McRae, 1997; see overview of English in Finland in Leppänen and Pahta, 2012, p. 146–147). In this sense, the concerted effort to introduce English language learning in the school curriculum reflects priorities connected to the education principles of the social welfare state (see section Definitions of Nordic Exceptionalism). At the same time, it would be faulty not to acknowledge that the concerted push toward English language learning was (and still is) a component part of globalization in relation to the outcome of World War II, with America emerging as a major player in global economics and politics, and Great Britain retaining, in part through the

continued use of the English language, at least some of its historical influence (Seargeant, 2012; Piller, 2016; Pennycook, 2017). Thus, the eventual outcome of widespread proficiency and use of English in the Nordic countries can be viewed as a mutually inclusive, if implicit, agreement: engagement in the modern global stage for the Nordic countries and a toehold via language and political and socioeconomic influence for the USA and Great Britain.

The eventual outcome of this marriage of intentions is that, as of the early 21st century, a majority of the population in each of the Nordic countries claims proficiency in English. A highly cited 2012 EU Barometer survey asked European Union citizens which languages they can speak well-enough to have a conversation (European Commission, 2012). The highest self-reports came from the UK and Ireland, with over 95 percent of respondents claiming they can have a conversation in English. The next highest reported score came from the Netherlands at 90 percent, followed by Denmark and Sweden at 86 percent, Austria at 73 percent and Finland at 70 percent[3]. While the results are based on self-reported proficiency, it is nonetheless of note that citizens of some EU nations, including some Nordic countries, lay claim to proficiency levels only a few percentage points below those of citizens of countries where English is the majority native language. Self-reports aside, demographic facts support the proficiency in English claimed by citizens of the Nordic countries. The English Language Proficiency Index (EPI), a for-profit English language skills company, measures the English language aptitude of people who are interested in knowing their ranking. Therefore, the measure is biased and non-representative—yet it nonetheless serves as an example of a global measure. Since the survey was first made available on the internet in 2012, ~1.7 million individuals have taken the exam. The survey results place countries in Northern Europe, including the Nordic countries and the Netherlands, in the same overall category as countries where English is a dominant mother tongue. In the EPI survey, the highest-ranking country was the Netherlands, followed by Denmark, Finland, Sweden and then Norway.

It is important to keep in mind that English is spoken as an additional, foreign language for the vast majority of people in the Nordic countries (excluding a tiny portion of the overall population which has English as a mother tongue). The European Parliament (2017) recommends that European citizens learn their mother tongue plus two foreign languages. In fact, in the Nordic countries at least, the common reality is that the overall population speaks English in addition to a domestic/national language, according to Eurobarometer data from 2006. This statistic is especially telling in the context of Finland, the only Nordic country that has two constitutional languages: Finnish and Swedish. According to Finnish law, Finnish pupils who speak Finnish as a mother tongue are required to learn Swedish

[1]Definitions of what constitutes the Nordic region of Europe vary. As sovereign political entities, the most common definition includes Denmark, Finland, Iceland, Norway and Sweden. Other classifications include the autonomous Faroe Islands (Denmark), Greenland (Denmark) and Åland (Finland). Still other definitions might include Russia, Estonia, Latvia and Lithuania, or other states in the northern parts of Europe.
[2]"Scandinavia" is a term often used to refer to the same geographical area as the Nordic. For the purposes of this article, Scandinavia is understood as comprising Denmark, Norway and Sweden.

[3]Iceland and Norway were not reported in the survey as they are not EU member states. It can be assumed that the overall language profile of Iceland and Norway patterns with other Nordic countries in their overall language profile + English. Notwithstanding, there are individual differences among Nordic countries (see e.g., Kristiansen, 2010).

at school, and pupils who speak Swedish as a mother tongue are required to learn Finnish at school (this is the general law; deviations exist according to region and language setting; Finnish Ministry of Justice, 2013). While in practice this means that Finnish citizens should be proficient in both of the two national languages, this is not the common reality: there is a marked decrease in widespread proficiency in Swedish (see e.g., Vaarala et al., 2017). According to the 2006 Eurobarometer report cited previously, 69 percent of the Finnish people polled report speaking only one language in addition to Finnish, and English is reported as being the main other language, spoken by 63 percent of the population. In Finland, 47 percent or respondents claimed to speak two languages in addition to their mother tongue. The main second additional language (after English) was Swedish, at 41 percent.

The general widespread proficiency in English in the Nordic countries has been explained through a number of factors. Common explanations, in no particular order, include: subtitled rather than dubbed English language television programs and films, the relatively small population base, and the fact that most citizens of the Nordic countries speak a Germanic language as their first language (see Peterson, 2019, p. 13–14). An additional explanation is the age of acquisition: Throughout the Nordic countries, formal classroom learning in English begins by Grade 3, with formal learning extensively boosted by informal learning outside the classroom (Norrby, 2015). A 2002 study (in Swedish; see English summary in Norrby, 2015) comparing the English skills of 9th grade students from eight European countries placed the Swedish and Norwegian students nearly consistently in the top tier across the four areas tested, with the Netherlands, Finland and Denmark occupying the middle tier and Spain and France the lower tier.

Each of these explanatory factors, however, can be countered, and therefore cannot be considered to have too much individual explanatory power. For example, many countries in the world have subtitled programs and films and yet do not exhibit high levels of language proficiency in the language of target media. Likewise, not all European countries with small populations exhibit overall high proficiency in English. In addition, Finnish and other languages spoken by Nordic populations (for example Meänkieli, the Sámi languages and Kven) are not Germanic languages (as the Scandinavian languages are), yet speakers of these Finno-Ugric languages are also able to achieve a high level of proficiency in English, despite speaking a genetically unrelated language as a mother tongue[4]. The most plausible explanation for the widespread proficiency is that there is a confluence of various factors, as described so far in this section. And, of course, an additional overarching factor that cannot be dismissed is the

enormous hegemonic social pressure to achieve high proficiency in English (Piller, 2016).

In Finland, like other Nordic countries, English as a classroom language starts in basic education, currently in grade 1[5]. English is the main (A1) foreign language for 89.9 percent of school children in Finland (Finnish National Agency for Education, 2019), especially for Finnish speaking children, who constitute the population majority. For Swedish-speaking children, who constitute about 5.3 percent of the total population of Finland, Finnish is the main A1. A large-scale survey on everyday language use in Finland, including the use of English, was conducted in the early 2000's, with the results published in 2011 (Leppänen et al., 2011). The survey, with responses from 1,495 people in Finland, demonstrated that the majority of people surveyed claim that English plays a significant role in their life; in fact, as many as 80 percent of the respondents said they encountered English in their everyday surroundings (Leppänen et al., 2011, p. 63).

DEFINITIONS OF NORDIC EXCEPTIONALISM

Definitions of what comprises the Nordic region vary, but many accounts include the five countries Denmark, Finland, Iceland, Norway and Sweden, all sovereign nations in Europe which lie above a 55 degree north latitude. While historically these nations have by no means constituted a consistently unified political bloc, there are nonetheless reasons to view them as sharing similarities in the current era (Vik et al., 2018). All of these countries are characterized by relatively small populations, especially in relation to the amount of geographic territory (most) occupy, and the governing system in each country is a social welfare democracy. Denmark, Norway and Sweden all have a constitutional government and a monarch as head of state. Iceland and Finland, as former territories of Denmark and Sweden, respectively, have a constitutional government with presidents as head of state. A governing union between Denmark and Norway ended in 1814. The sovereign nation now known as Finland was an eastern territory of Sweden for several centuries, until 1807. Iceland became a sovereign nation in 1918 but shared a monarch with Denmark until 1944.

As pointed out in the previous section, there are two main language families represented among the populations of the Nordic countries—although clearly there are many more languages among the total population due not only to migration but also with respect to long-standing minority groups, such as speakers of Russian, Romani and Tatar (Latomaa and Nuolijärvi, 2002). The most heavily represented languages population-wise are the Scandinavian languages, a subset of the Indo-European/North Germanic language group, further divided into Old East Norse (Swedish and Danish) and Old West Norse (Icelandic and Norwegian). The second most represented are Balto-Finnic languages, including Finnish and the Sámi languages. Obviously, linguistic boundaries do not coincide with national borders. Finland, for example, can be considered part of

---

[4]While, overall, Finnish speakers have achieved a high level of proficiency in English according to the measures cited in the article, it is important to note that the overall level of proficiency reported is lower than other Nordic countries where a Germanic/Scandinavian language is the main language. In addition, research in Finland shows that Swedish-speaking Finns exhibit a lower level of nonstandard features in their English than Finnish-speaking Finns, lending credence to the argument that L1 could be a factor (see Meriläinen, 2021 – and thank you to an anonymous reviewer for pointing this out).

[5]Grade 1 since 2020; prior to 2020, the A1 language was introduced in Grade 3. The numbers reported here are from 2017.

the Scandinavian speaking area (Norrby, 2015) due to its shared history with Sweden, the fact that Swedish is a constitutional language, and part of its population has Swedish as mother tongue. Likewise, there are Balto-Finnic-speaking populations outside of Finland, for example in the northern regions of Norway, Sweden and Russia, as well as longstanding heritage language speakers throughout Sweden (see e.g., Sundberg, 2013 for more information).

In terms of alliances and official associations, three Nordic countries, Denmark, Finland and Sweden, are in the European Union (EU), but all five countries, including Iceland and Norway, are part of the European Economic Area (EEA). In recent decades (as well as historically; the Kalmar Union existed from 1397 to 1523), the Nordic countries have purposefully enhanced their cooperation and cohesiveness through a shared Nordic cooperation[6]

It is important to establish this shared set of both real and potentially imagined factors that combine the nations of the Nordic area into a logical and self-evident bloc. For the purposes of the arguments presented in this article, this is a crucial foundation in viewing the concept known as *Nordic exceptionalism*, a term that seems to have been circulating roughly since the early 2000s[7].

Exceptionalism is a collective (often self-held) belief that a group or community, for example a nation state, possesses inherent characteristics that distinguishes it and makes it more special than other groups or nations. The term *exceptionalism* has often been used in conjunction with ideologies of the United States: *American exceptionalism,* which has been defined as "a political doctrine as well as a regulatory fantasy that enabled U.S. citizens to define, support, and defend the U.S. national identity" (Pease, 2009, p. 11). Further fuel for America's sense of exceptionalism draws from its dominance on the world stage after World War II, its historical lack of a feudal system, and its unique historical background (among other factors; see e.g., Fredrickson, 1995; Pease, 2009). It is crucial to note that a shared sense of exceptionalism makes it possible for the groups who consider themselves "exceptional" to avoid responsibility or blame due to the overriding ideology that they are a special group to which the "usual" rules do not apply, thereby allowing the group to eschew culpability or avoid compensatory action for systemic racism, colonization, or genocide, for example (for an overview on the context of Finland, for example, see Hoegaerts et al., 2022).

Within the context of the Nordic countries, exceptionalism is treated, at least for the purposes of this article, as emerging from two distinct but overlapping directions. One set of ideologies is borne out of the current era, relating largely to the outcome of the social welfare model and discourse surrounding the "success" of the Nordic countries with respect to social equity. The second

set of ideologies relates to colonial exceptionalism. Within the Nordic context, colonial exceptionalism is perpetuated through the belief that the Nordic countries were not actively involved in the colonization of the Americas, Asia and Africa to the extent or in the same manner as other European nations, thereby affording an overall notion that Nordic countries are in some ways "innocent" (see Keskinen et al., 2019). In some Nordic countries, including Finland, such ideologies are further enhanced by the historical absence of a monarchy and an elite ruling class (compare to exceptionalism in the United States and the historical lack of a feudal system).

The Social Welfare Model in the Nordic Countries

While the typical system of governance in much of Europe is the social welfare model, a system dependent on taxation of citizens for the redistribution of goods and services, the Nordic countries stand out as being particularly well-known for this model. In fact, many would consider the Nordic countries to be exemplars of the social welfare model (Hilson, 2008). For most people in the Nordic countries, education, health care and childcare are provided by the state, either free of charge or heavily subsidized (Pratt and Eriksson, 2013, p. 66). Unlike in countries that do not utilize the social welfare model, or enact it to a lesser extent, these services are not just available for the lower socioeconomic classes, but for everyone. The availability of these services for everyone, regardless of socioeconomic class, is in fact a key equalizing component of the model[8]. Assistance is also available for certain non-citizens, for example new migrants who are not yet employed.

Citizens in the Nordic countries can be taxed as much as 80 or 90 percent, depending on income. The average taxation rate in Finland is about 40 percent, compared to a global 34 percent [Organisation for Economic Co-operation and Development (OECD), 2021]. Sweden and Denmark were the first Nordic countries to adopt this particular social welfare model relying on taxation, with Finland adopting the system later on (for more, see Kangas and Palme, 2005; Kvist et al., 2012). Social services expanded from the post-World War II period into the 1960's, peaking in the 1970's (ibid) concurrent, incidentally, with changes to the education system, including the introduction and subsequent dominance of English as a foreign language in the education system.

Explanations for the "success" of the social welfare system in the Nordic countries vary. It has been pointed out, for example, that the Nordic countries "went on to reproduce their already existing value systems: the emphasis on moderation, egalitarianism and uniformity [...]; the emphasis on individual responsibility" (Pratt and Eriksson, 2013, p. 66) in many ways carried on functionalities and ethos of the region's most dominant religion, Lutheranism. The outcome is that the Nordic countries are those in Europe to demonstrate most precisely the

---

[6]Available online at: https://www.norden.org/en/information/official-nordic-co-operation.

[7]Searches for the term *Nordic exceptionalism* yielded zero hits on corpora such as the BNC and COCA at the time this article was being written. However, research on the topic indicates that the term began circulating in academic circles in perhaps the early 2000s; (see, e.g. Delhey and Newton, 2005).

[8]An anonymous reviewer points out another important equalizing factor: the "freedom to roam" laws, entitling every person to access nature even on privately held lands – for example hiking, fishing, and picking mushrooms and berries. This freedom to roam law exists in all of the Nordic countries except Denmark.

principles of the social welfare system, and, in relation, they are generally regarded as having made the greatest efforts to curb social inequity (Kvist et al., 2012, p. 5). The properties play a key role, for example, in the selection of the Nordic countries as the "happiest," according to rankings reported by Cambridge, Gallup, and the World Happiness Report.

In 2012, the United Nations launched its first World Happiness Report (Helliwell et al., 2012), a ranking carried out by several group and independent experts. Denmark was ranked in the number one position in the report's 1st year, with Finland and Norway in second and third positions, respectively. In the ensuing years, the Nordic countries have been ranked consistently in the top positions, with Finland being ranked first for the past 4 years, 2018 to 2021. The ranking is based on scores across several different areas, including gross domestic product per capita, social support, healthy life expectancy, freedom to make one's own life choices, generosity of the general population, and perceptions of internal and external corruption levels (Helliwell et al., 2021). An interesting twist to the World Happiness Reports is the extent to which Anglo-centric notions of "happiness" map—or rather, don't map—onto to existing notions in the Nordic countries (Levisen, 2012, 2014). A range of reactions, from the academic to the journalistic (e.g., Savolainen, 2021) venture that "happiness" in the Nordic is more accurately described as being content with what one has, a concept that relates back to the overall tendencies described previously. Levisen (2012) demonstrates compelling evidence that the Anglo concept of "happiness" does not accurately translate into Danish, nor other languages, rendering it a vague or even misleading measure. Indeed, backlash commentary about the "happiness" rankings tends to refer to the relatively high rate of depression and suicide in Nordic countries, compounded by the darkness and harsh weather that exists much of the time (e.g., The Guardian, 2018).

As a summary, then, the "success" of the social welfare model in the Nordic countries has fed into numerous reports and discourse which actively indulge ideologies about Nordic exceptionalism, contributing to a sense of pride and positive branding (Levisen, 2012). This is a contemporary manifestation of exceptionalism that, prior to the 1970's, could not have existed in what were, for the most part, relatively poor, marginalized and peripheral populations. At the same time, it should be noted that there is a decided backlash and what could even be considered a counter ideology to the "happiness" reports.

Colonialism and Exceptionalism

In the current era, a great deal of research has treated issues of colonialism, migration and language in the Nordic countries (Loftsdóttir and Jensen, 2012; Aaltonen and Sivonen, 2019; Keskinen, 2019; Keskinen et al., 2019; Kujala, 2019; Ranta and Kanninen, 2019), highlighting the tension between a sense of exemption—also termed *exceptionalism*—on the one hand, and a history of colonialism as well as contemporary racism on the other. For example, Finland, the "happiest" country for 4 years running, has also been named the most racist country out of 12 EU countries studied (European Union Agency for Fundamental Rights, 2018).

Finland and the other Nordic countries were never colonized by Britain or the United States, although Iceland was occupied by British, Canadian and US troops during World War II. Some of the Nordic countries, namely Denmark and Sweden, were involved in the European colonization of the Americas and Africa, with the other Nordic countries engaging in involvement through their governing affiliation with Sweden and Denmark, or through religion and missionary work (see Keskinen et al., 2019). In addition, there are both historical and contemporary examples of colonialism in the Nordic region involving, for example, Danish rule of Norway and parts of Sweden, Swedish rule of Norway and Finland, and Danish rule of Iceland, not to neglect the ongoing Danish control of Greenland. A discussion of colonialism in the Nordic countries would be remiss not to highlight the exceptional mistreatment of Sámi territory and people, and the attempted erasure/integration of the indigenous languages and cultures (see e.g., Kuokkanen, 2020), as well as the segregation and attempted erasure of linguistic groups such as the Kven in Norway (Lane, 2016).

A comprehensive description of many of these colonial relationships can be found in a volume by Loftsdóttir and Jensen (2012), who also use the term *exceptionalism* to refer to a pattern of exemption or non-culpability. For example, drawing on discourse about "happy" and "peace-loving" people (Loftsdóttir and Jensen, 2012, p. 2) note that exceptionalism can refer to "current forms of internationalism" (ibid., also addressed in section The Social Welfare Model in the Nordic Countries of this article), as well as nation branding and contemporary activities since the 1970's that pose anachronistic conflicts with a history of domination and conquest during the colonial period.

Such notions are perhaps especially interesting when applied to the context of Finland. Historically, a common ideology among Finnish people has been that they are an oppressed people, even victims (see Hoegaerts et al., 2022; see also Keskinen, 2019). Such an ideology is the outcome of several historical factors, including the fact that Finland did not gain independence until 1917, after more than 100 years as a Grand Duchy of Russia and hundreds of years as a territory of Sweden before that. Finland has been a relatively poor, agrarian country, with what was for much of its history considered a marginalized vernacular, Finnish, along with a majority population who were in some accounts considered racially inferior (ibid.). It was only in recent decades that Finland has emerged on the world stage as a competitive force. These factors may contribute further to a sense of exceptionalism: it can be difficult to critically observe inequity and oppression when an overriding ideology is that of succeeding in spite of being a victim.

THE ENGLISH LANGUAGE AND COLONIALITY

English is not and never has been an empty vessel of a language, devoid of historical and cultural baggage. On the contrary, the language itself is permeated with and has been created through conquest and exploitation. A foremost fact that should underlie any discussion of the "success" of English is exactly how English

came to occupy this rarified status (see Saraceni, 2019). While in the current day and age it is not uncommon to hear arguments that English is a "neutral" choice (Lemberg, 2018) this is in fact not a straightforward assessment. Rather, as with the discussion of exceptionalism in the Nordic context, it is an assessment that clouds over the fact that present-day widespread use and accessibility of English is entrenched in historical injustices and domination (see also Pennycook, 2017). Furthermore, it should not be overlooked that the English language norms that make their way into foreign language textbooks and guides in English-speaking settings around the world are based on the standardized forms of English, which are directly linked to the English language use of elite, white, upper-class speakers (see e.g., Lippi-Green, 2012; Peterson, 2019, p. 17–27; Peterson, 2020; see also Pennycook, 2017).

These ideals of English are ideals of English are so successfully transferred to foreign language settings that so far it is unattested for English to undergo stabilization in the same way it has, for example, in some post-colonial settings (Buschfeld and Kautzsch, 2017). The prescriptive transference of norms and ideologies has been attributed to a number of factors, including not only ideologies present in language teaching materials, but further reinforcement of exonormative ideals through English language media (ibid 118). What this means in practice is that places with no direct colonial history with Great Britain or the United States adopt the English language ideologies that are established by the norm-providing countries, although ostensibly it should be possible to not do so. Of the numerous different perspectives on the development of English as a global language, some posit that it would actually be highly unlikely for English to be anything other than exonormative (that is, norm-following, to use terminology from Kachru's classic 1982 delineation) in foreign language settings, in large part because it is an additional language rather than a widespread first language. Therefore, it is mostly a learned language rather than a language acquired from intergenerational transmission or a speech community. In other words, language settings such as Finland do not create the right conditions for endormativity and might even, in fact, preclude this possibility, making the adherence to external norms of English a necessity (see Leppänen and Nikula, 2008, p. 21)[9].

COMBINING THE THREADS: EXCEPTIONALISM AND THE ENGLISH LANGUAGE IN FINLAND

So far the discussion in this article has concentrated on first establishing shared characteristics in the Nordic countries that led to a phenomenon best described as Nordic exceptionalism, followed by an overview of coloniality in relation to the English language itself. For the remainder of the article, these broader notions are applied to a much narrower focal area: Finland. The purpose of this exercise in effect demonstrates that a sense of

exceptionalism with regard to the use of English runs parallel with a sense of exceptionalism in other areas.

In a discussion of English use at the level of the individual speaker, I have argued (Peterson, 2019, p. 76) that a foreign-language user of English has an opportunity (aptitude notwithstanding) of stylistic choice and identity through English—for example, by espousing a New York City accent—that first and second language speakers might not possibly have access to, in part for ethical reasons. The same level of choice and freedom does not likely apply at the national level; for example, when it comes to the teaching of English, basic accessibility of materials and pedagogical practicality point toward the adoption of a standardized, prescriptive variety of English (Young and Walsh, 2010). A possibility that can be espoused through a national curriculum, however, is that of tolerance and education about the rich expanse of varieties of English, as well as imbuing learners with an understanding of the sociolinguistic underpinnings and relative esteem placed on these varieties. Put in another way: does increased proficiency in English cause a population to become more tolerant and accepting of the multiple ways of using English, or does it induce even further instantiation of prescriptivism and coloniality of language use? These are the notions that are explored in the remainder of the article. The exploration concentrates on three main areas: social class, language attitudes about English speakers, and social exclusion.

Social Class and English

An outcome of the ideology of equality in the Nordic region, in this discussion Finland in particular, is that it is not a class-based society (see e.g., Erola, 2010). With regard to spoken English, particularly in Great Britain, it is well-established that accent is a strong indicator of social class (see Hughes et al., 2012; Sharma et al., 2019). Here, the term *accent* refers specifically to pronunciation and prosodic features. In Finland, accents of spoken Finnish are generally more perceived as indexing region than socioeconomic class, although clearly these factors can be closely connected (Nuolijärvi and Vaattovaara, 2011)[10]. When it comes to English accents and social class, however, Finnish people nonetheless can exhibit a high degree of awareness of the social capital of certain accents over others (see Bourdieu, 1977), as demonstrated here.

For example, Peterson (2022), presents a journalistic sample obtained from an article in Finland's major daily newspaper, *Helsingin Sanomat*. The newspaper article, which describes the spoken English language skills of then-Minister of Finance, Katri Kulmuni, begins with the headline and subheading "Katri Kulmuni's delightful British English dropped jaws on social media—we asked linguists if she would be an upper- or lower-class Brit" (*Helsingin Sanomat* 2020, January; author's own translation). The article goes on to describe how an assessment from various linguists confirmed that the Finnish Minister of Finance, a native speaker of Finnish and a foreign language speaker of English, spoke English like an upper-class [sic]

---

[9]Thanks to an external reviewer for highlighting this perspective and for the example from the literature.

[10]As this article focuses on English in Finland, second-language accents of Finnish are not discussed. The topic of migrants in Finland is addressed later in the article.

Brit—which she asserted she learned in part from watching the television series Emmerdale. The fact that a report of this nature was deemed newsworthy in a daily newspaper in Finland merits further inspection. First, it demonstrates a high-level of awareness and, whether intentional or not, complicit compliance with socioeconomic class divisions and their relation to British English accents. This fact is further emphasized by its presence in a major daily newspaper: while it impossible to validate, it is unlikely the story would have been deemed newsworthy if the minister in question spoke working class British English (an unlikely but still plausible scenario), or, indeed, "Finnish" English.

The reactions to this politician's way of speaking English show a high level of regard for a prestigious variety of British English, a level of awareness which is directed toward an external model of English (as discussed in section The English Language and Coloniality). In comparison, there is evidence of a low level of regard for an internally influenced way of speaking English, or in other words English that is perceived as exhibiting strong influence from L1 Finnish. Indeed, there are a few names, all derogatory, for English used by Finns that is perceived as exhibiting strong influence from Finnish, including *rallienglanti* and *tankeroenglanti.* The former term comes from Finnish rally competitive drivers and their use of English in media interviews. The latter term is associated with a former Finnish prime minister who was reported to mispronounce the English word *dangerous,* instead uttering something that was heard as "tankeros." While they can hardly be considered representative of everyday Finns and their attitudes toward English, it is nonetheless noteworthy that an informal survey of English majors at the University of Helsinki shows a preference to speak with a "British" accent if possible. None of the students polled expressed a desire to speak English with a "Finnish" accent. English majors are likely a biased group of English language users, and in contrast, research of workplace English in Finland indicates a high level of negotiability and context sensitivity to the use of language, including English pronunciation. That is, there seems to be a level of acceptance in practice of Finnish-accented English (Oksaharju, 2021), although the examples used here show that politicians, no doubt due to their public status, can be held to a higher level of scrutiny.

Moving beyond the level of pronunciation, it is telling to note the relationship between the English language—indeed, certain foreign languages in general—and social class. Historically, knowledge of foreign languages has been a marker of higher socioeconomic class and intelligentsia in Finland. With the Finnish nationalist movement, starting at the end of the 19th century "authors, poets, writers and journalists … resembled the archetype of a French intellectual. They were academically educated, well-traveled individuals, skilled in languages— virtual renaissance figures, who participated in artistic circles and were interested in various social, political and cultural issues. The resemblance to France was not coincidental" (Kortti, 2014, p. 7). An overview of the history of the study of English in Finland (Pahta, 2008) demonstrates a clear connection between the characteristics described with the intelligentsia of the nationalist movement and the establishment of English as an academic field of study in Finland. In the current era, the relationship of

foreign language access and social class continues. For example, compared to municipality-owned schools, the number of foreign languages learned in private and state-owned schools is higher, 41 percent compared to 20 percent. Research shows that students from higher socio-economic rankings are more likely to study multiple foreign languages other than English, while students from poorer and more rural areas are likely to have access only to English (Finnish National Agency for Education, 2019; see also Vaarala et al., 2017).

Whose English Is "Best"?

A topic related to social class and pronunciation of English is the social hierarchy imposed on varieties of English. English, of course, is the very definition of a pluricentric language (Clyne, 1992); given its colonial history and subsequent emergence as a global language and lingua franca (Crystal, 2003)—it is a language with many homes and different kinds of speakers. An issue raised earlier in this article questions if populations who use English as a foreign language are inclined to be relatively accepting of different varieties of English, or rather if they enact language attitudes mirroring those of inner circle, aka native-speaker settings (see Peterson, 2019). An openness toward varieties of English would be in keeping with ideologies of equality which are for many perceived to be part of Nordic exceptionalism. Indeed, tolerance toward variation of domestic languages in the Nordic countries is a hallmark of overall equality (but see, for example, Røyneland, 2009). It is therefore interesting to observe if the tolerance and acceptance of varieties of English is in evidence, or rather if externally modeled attitudes against certain varieties of English manifest.

As early as the 1990's, research on immigration to Finland highlighted the linguistic allowances afforded to English speakers from certain countries, for example the United States (Latomaa, 1998). English-speaking immigrants from inner-circle settings (Kachru, 1982) to Finland are stereotyped and indeed are demonstrated to be in a privileged position, hinging in large part on their language capital (Latomaa, 1998; Latomaa and Nuolijärvi, 2002; Koskela, 2020). What this means in practice is that an English-speaking migrant, especially one from a rich nation, is in a more amenable social position than an immigrant from a poorer nation, especially one who does not speak English.

These characteristics were reported in a recent PhD thesis (Koskela, 2020) on migrant "elites" to Finland. An example from the researcher's fieldnotes described an incident when a Finnish migrant from Kenya, a Black man, needed to defend himself against the unwanted verbal abuse of two (intoxicated) Finnish men who approached him on the street in Helsinki. The Kenyan man, who spoke good Finnish, nonetheless chose to respond to the two harassers in English, thereby switching not only the language of the exchange but also the content. Through speaking English, he was able to present himself as an educated, skilled migrant—and he was thus able to distinguish himself from other Black migrants, who are often associated with refugee status (Koskela, 2020). In fact, when asked by the researcher about his choice to switch to English during the exchange, the man explained, "'if you're black and you speak Finnish, it's more likely that you came as a refugee. If the person speaks fluent Finnish,

the person has gone into the integration plan" (Koskela, 2020, p. 6). With this observation, the person in question refers to the language skills programs that are typically made available for refugees who arrive as asylum seekers, with the aim to help them integrate into the work force and to participate more fully in Finnish society. For various reasons, clearly including race but also language, migrants from African and Asian countries to the Nordic region have been less successful at gaining employment than those from other regions (Gerdes and Wadensjö, 2012, p. 192). Overall, these systematic disadvantages lead to increased levels of poverty, crime and lack of access to public goods and services.

While there do not appear to be scientific studies of this phenomenon, there is evidence in the current era that some workplaces, even English-speaking workplaces, prefer to hire English-speaking Finnish people rather than English-speaking migrants. If demonstrable, this tendency would point toward an acceptance of Finnish-English on par with that of a native speaker from an inner circle setting. An example from my own field notes (personal communication, November 12, 2021) serves as an illustration. In an interview from an ongoing project on English in the workplace, a migrant, highly skilled worker in the field of professional communication explained that they applied for a job at the headquarters of a multinational Finnish-based company. The job description they responded to called for applicants who were native speakers of English, specifying that the work would be carried out 100 percent in English–as English was the workplace language for the company headquarters, in Finland. The job applicant was a native speaker of US English and did not speak Finnish. The person was highly qualified with many years of experience in the relevant field. However, the applicant was informed by a human resources director that the company was going to hire a Finnish person who could speak Finnish as the social language of the company and working group—despite the fact that the job description explicitly requested a native speaker of English who would work totally in English.

The previous examples in this section indicate that there is high social value placed on "native"-like accents produced by Finnish people when they speak English. The latter examples indicate, at the same time, a preference for a Finnish speaker of English over a native speaker of English—even one from an inner-circle setting—for certain workplace settings. A recent study by Peterson and Hall (2017) likewise demonstrates a tendency to place Finnish L2 English (and other Nordic Englishes) on par with inner-circle Englishes within the context of institutions of higher education. This study observes data from Nordic universities' admission requirements for international MA programs in English. Higher education in Finland is free of charge for EU citizens and EEA member states, but since August 1, 2017, tuition fees have been in place for students who are not EU or EEA member states citizens (Finnish Ministry of Education and Cultur, 2020). Individual universities and university programs differ with regard to not only their fee rates but also their admissions criteria, which extends to language requirements, as well. In recent years, there has been

an upsurge in the number of English language medium programs in institutions of higher education in the Nordic countries, as detailed, for example in Hultgren et al. (2014). In 2020, English medium programs in Finnish Institutes of Higher Education received a record number of applications from foreign applicants, 8000 – more than 2000 more than received the previous year (Finnish National Agency for Education, 2021). The language medium of education, as well as the content, is indicative of the changing role and related tensions of achieving global competitiveness while at the same time serving the needs of the nation and welfare state (Hultgren et al., 2014; Saarinen and Taalas, 2017; Saarinen, 2020).

An analysis of English language requirements from 25 universities in five Nordic countries (Peterson and Hall, 2017) demonstrated that applicants to English medium programs from six English speaking countries—Australia, the Republic of Ireland, English-speaking Canada, the USA and the UK— are exempt from having to submit proof of proficiency in English, for example through a standardized test such as the TOEFL. While in one way this allowance seems straightforward, it nonetheless draws a concerning line dividing relatively more privileged English speakers from those who are less privileged. While these six English-speaking nations were uniformly exempt across settings, exemptions for other English-speaking regions varied widely. For example, in Icelandic universities, citizens of English-speaking countries throughout the Americas and the Caribbean are exempt from proving proficiency in English, whereas universities in Sweden offer certain kinds of exemptions for citizens of India, Pakistan and 17 different English-speaking African nations. Narrowing the focus to Finland, the language requirements of the University of Helsinki—the highest ranking and biggest university in Finland—specify that English-language exemptions are offered to applicants who finished upper secondary school in the Nordic Countries, Australia, Canada, Ireland, New Zealand, the UK or the USA. In addition, exemptions are made for applicants who have a university degree in English from an EU/EEA country or Switzerland, Australia, Canada, New Zealand, the UK or the USA. While various other exemptions are made—for example, a diploma from an International Baccalaureate program—the details already supplied merit further inspection. Perhaps most noteworthy is the fact that across Nordic universities, there are two kinds of English speakers who are uniformly exempt from demonstrating language proficiency through standardized tests: assumed native speakers who come from so-called inner-circle settings, and foreign language speakers of English from Nordic countries (and, in some instances, from EU/EEA countries).

From a purely practical standpoint, it can be considered reasonable that Nordic universities would favor applicants from their own citizenship, especially considering the role of universities in contributing to nationhood and also Nordic cooperation. English, being the language currently most used in higher education, plays an obvious role in the global viability of any university. What is noteworthy about this study, however, is the evidence it supplies about Nordic countries not only mimicking but reinforcing inequity with regard to varieties of

English. It might be argued that this is a price to pay for being competitive on the global front, but at the same time a more critical stance would acknowledge the benefits of participating in coloniality through the use of English.

Exclusion From English

To date, the only large-scale survey on the use of English in Finland (Leppänen et al., 2011; discussed in section English Language Proficiency in the Nordic Countries) was conducted nearly 20 years ago. A similar survey today would likely gain at least some different results, considering the wide reach of English-based social media even to peripheral areas in recent history (see e.g., Laitinen et al., 2020). Considering the survey as an output of its time, it is of note to recognize that it demonstrated a clear divide in terms of region and age. English tends to be a more a part of the everyday experience of people in urban regions of Finland compared to rural regions. Given the relatively recent history of widespread education in English as a foreign language in Finland, it is not surprising that the survey showed that many Finns aged 45 and above (at the time of the survey) had not studied English at all, and that older people rated their own English skills considerably lower than younger people (Leppänen et al., 2011, p. 94, 97).

In the context of Finland, the main urban region is the capital city area, Helsinki, geographically located in the southern portion of the nation. The greater urban Helsinki area is home to some 1 million inhabitants. The overall population of Finland is 5.53 million. In addition, there are several relatively larger municipalities serving as local urban centers; for example, Oulu, population 200,000, situated in the west central region of Finland (see Vaattovaara and Peterson, 2019). Research on the use of English language borrowings into Finnish has demonstrated that such borrowing are more associated with urban speech, and survey data also demonstrates that urban dwellers are the most likely to claim using the English language borrowings tested. For example, research on the borrowing *pliis* "please" in Finnish indicated that it is indexical of an urban style (Peterson and Vaattovaara, 2014), especially younger, urban women. Likewise, a survey investigation of English-language swearword borrowings into Finnish demonstrated that many of the respondents associate English swearword variants, for example *damn,* as more regionally neutral, whereas the Finnish swearword variant, *perkele* "damn" was more associated with rural (and male) speakers. Note, however, that while open (written) comments from the respondents supported the notion of an urban/rural division of use, regression analyses of the claims of usage of *damn* and *perkele,* modeled against demographic features, did not show significant differences for region. In fact, age was the only factor consistently significant in the regression analyses of *damn* in comparison to Finnish variants *perkele* and the mild swear word *hitto* "darn." Other demographic features, such as region, level of education and gender were not consistently significant across the models.

Such findings point toward intersectionality, but also toward a reality in which certain segments of the overall population, including the elderly and those who live in rural areas, do not have the exposure to English that would ensure their inclusion in many aspects of contemporary linguistic life in Finland. Such a reality is detailed, for example, in an ethnographic investigation of an elderly married couple, a man and woman, who live in rural Finland (Pitkänen-Huhta and Hujo, 2012). The elderly woman in the study, whom the researchers call Aino, relates that when she and her husband go into town, they "bump into" English words that come up, words they do not understand. The same couple also reflect on the fact that it is more likely to need English in bigger towns, where there are more foreigners. At the same time, however, the couple reveal that they have learned a few English expressions from television, for example "I love you."

CONCLUSIONS

The social aspects explored in the article—social class and its relation to the English language, attitudes toward varieties of English and social divisions—are certainly not specific to Finland, nor is it a realistic or an intended aim to point the finger at Finland as a particularly problematic case. While a cross-setting analysis is beyond the scope of this article, it can be assumed with a high level of certainty that such divisions and inequities with regard to English can and do exist in other settings. The critical stance taken here is not intended to push Finland into a critical or damning light, but rather, as the author's academic home and research setting, Finland has been utilized as a case study to explore the relationship of Nordic exceptionalism and the use of English.

Such an exploration is revealing for several reasons. For many people around the world, the Nordic states, including Finland, serve as highly regarded models of social equity, education, and other desired societal affordances. In various global rankings, these countries top the list of "most equal" in the world. An outcome of this level of recognition is a phenomenon termed *Nordic exceptionalism.* The same term is used in application to a historical disavowal of colonialism or even colonial complicitness—in essence, an assumed innocence. These notions are not entirely borne out in reality, neither in the past nor in the present. The role of English in itself becomes a telling tool for further inspection of these properties. That is, contemporary English is the outcome of the colonial history of Great Britain, and the vastness and varieties of English hold testament to the fact that not all Englishes hold equal prestige across settings. Indeed, certain varieties of English are demonstrated to be associated with grave social injustices levied against people who speak them, for example in the workplace, in schools and in courts of law (see Peterson, 2019).

A question posed earlier in this article is reintroduced: does increased proficiency in English cause a population to become more tolerant and accepting of the multiple ways of using English, or does it induce even further instantiation of prescriptivism and coloniality of language use? These possibilities are of unique application to the context of the Nordic countries. As established earlier in this article, the populations of the Nordic countries are among those in the world that exhibit some of the highest levels of overall proficiency in English. At the same

time, these are also some of the countries in the world which are most regaled and recognized for their efforts to promote equity and equal opportunity to social benefits such as education. What then, about equality and English? The evidence presented in this article, stemming from the setting of Finland, suggests that equity and the use of English are not, unfortunately, a straightforward pairing. However, to address this imbalance, some policy makers and educators in Finland are actively working to take into account the special properties of the English language today, namely its wide variability and multiple uses, such as English as a lingua franca. For example, the most recent version of the New National Curriculum for basic studies (Finnish National Board of Education, 2016) calls for increased exposure to varieties of English.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Aaltonen, J., and Sivonen, S. (2019). *Orjia ja isäntiä – Ruotsalais-suomalainen siirtomaaherruus Karibialla.* Helsinki: Into Kustannus.

Bourdieu, P. (1977). The economics of linguistic exchanges. *Soc. Sci. Inform.* 16, 645–668. doi: 10.1177/053901847701600601

Buschfeld, S., and Kautzsch, A. (2017). Towards an integrated approach to postcolonial and non-postcolonial Englishes: Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Eng.* 36, 104–126. doi: 10.1111/weng.12203

Clyne, M. G. (1992). *Pluricentric Languages: Differing Norms in Different Nations, Contributions to the Sociology of Language.* New York, NY: Mouton de Gruyter. doi: 10.1515/9783110888140

Crystal, D. (2003). *English as a Global Language.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511486999

Delhey, J., and Newton, K. (2005). Predicting cross-national levels of social trust: global pattern or nordic exceptionalism? *Eur. Sociol. Rev.* 21, 311–327. doi: 10.1093/esr/jci022

Erola, J. (2010). *Luokaton Suomi? Yhteiskuntaluokat 2000-luvun Suomessa.* [Classless Finland? Social classes in 21st century Finland] Helsinki: Gaudeamus.

European Commission (2012). *Special Eurobarometer 386: Europeans and Their Languages.* Available online at: http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf/

European Parliament (2017). *Language Policy. Fact Sheets on the European Union.* Available online at: http://www.europarl.europa.eu/factsheets/en/sheet/142/language-policy.

European Union Agency for Fundamental Rights (2018). *Being Black in the EU. Second European Union Minorities and Disctimination Survey – Summary.* Available online at: https://fra.europa.eu/en/publication/2019/being-black-eu-summary

Finnish Ministry of Education and Culture (2020) *Basic Education.* Available online at: https://minedu.fi/en/basic-education

Finnish Ministry of Justice (2013). *Report of the Government on the Application of Language Legislation.* Available online at: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/80741/englanti_web2013.pdf?sequence=1&isAllowed=y

Finnish National Agency for Education (2019). *What Languages Do Pupils Study in Basic Education?* Available online at: https://www.oph.fi/sites/default/files/documents/factsexpress1c_2019.pdf

Finnish National Agency for Education. (2021). *Number of Applicants by Field, First Joint Application to Higher Education.* Available online at: https://www.oph.fi/sites/default/files/documents/kevaan-1.-yhteishaun-hakijamaarat_2020_en_11_2_2020_0.pdf (accessed February 18, 2022).

Finnish National Board of Education. (2016). *National Core Curriculum for Basic Education 2014.* Helsinki: Finnish National Board of Education.

Fredrickson, G. M. (1995). From exceptionalism to variability: recent developments in cross-national comparative history. *J. Am. History* 82, 587–604. doi: 10.2307/2082188

Gerdes, C., and Wadensjö, E. (2012). "Is immigration challenging the economic sustainability of the Nordic welfare model?" in *Changing Social Equality: The Nordic Welfare Model in the 21st Century,* eds J. Kvist, J. Fritzell, B. Hvinden, and O. Kangas (Bristol and Chicago: Policy Press). doi: 10.1332/policypress/9781847426604.003.0009

Helliwell, J. F., Richard, L., and Jeffrey, S. (2012). *World Happiness Report.* New York, NY: UN Sustainable Development Solutions Network.

Helliwell, J. F., Richard, L., Jeffrey, S., and Jan-Emmanuel, N. (2021). *World Happiness Report.* New York, NY: Sustainable Development Solutions Network.

Hilson, M. (2008). *The Nordic Model Since 1945.* London: Reaktion Books.

Hoegaerts, J., Liimatainen, T., Peterson, E., and Hekanaho, L. (2022). *Finnishness, Whiteness, and Coloniality.* Helsinki: Helsinki University Press.

Hughes, A., Trudgill, P., and Watt, D. J. L. (2012). *English Accents & Dialects: An Introduction to Social and Regional Varieties of English in the British Isles (Fifth edition).* London: Hodder Education. doi: 10.4324/9780203784440

Hultgren, A. K., Gregersen, F., and Thøgersen, J. (2014). *English in Nordic Universities: Ideologies and Practices.* John Benjamins Publishing Company. doi: 10.1075/wlp.5

Kachru, B. B. (1982). *The Other Tongue: English Across Cultures.* Chicago: University of Illinois Press.

Kangas, O., and Palme, J. (2005). "coming late — catching up: the formation of a 'nordic model.'" in *Social Policy and Economic Development in the Nordic Countries,* eds O. Kangas and J. Palme (Palgrave Macmillan). doi: 10.1057/9780230523500_2

Keskinen, S. (2019). Intra-nordic differences, colonial/racial histories and national narratives: rewriting finnish history. *Scand. Stud.* 91, 163–181. doi: 10.5406/scanstud.91.1-2.0163

Keskinen, S., Skaptadóttir, U. D., and Toivanen, M. (2019). *Undoing Homogeneity in the Nordic Region: Migration, Difference, and the Politics of Solidarity* (1st ed.). London: Routledge. doi: 10.4324/9781315122328

Kortti, J. (2014). Media, the elite and modernity - defining the modern among the finnish cultural intelligentsia in the twentieth century. *Int. J. History Cult. Modernity*, 2, 1–24. doi: 10.5117/HCM2014.1.KORT

Koskela, K. (2020). *Negotiating the 'Migrant Elite': Boundary Making and Social Identities Among Skilled Migrants in Finland.* PhD thesis, University of Helsinki.

Kristiansen, T. (2010). Conscious and subconscious attitudes towards English influence in the Nordic countries: Evidence for two levels of language ideology. *Int. J. Sociol. Lang.* 2010:204. doi: 10.1515/ijsl.2010.031

Kujala, A. (2019). *Kivenmurskaajat. Kolonialismin historia.* Jyväskylä: Atena Kustannus.

Kuokkanen, R. (2020). Reconciliation as a threat or structural change? The truth and reconciliation process and settler colonial policy making in Finland. *Human Rights Rev.* 21, 293–312. doi: 10.1007/s12142-020-00594-x

Kvist, J., Fritzell, J., Hvinden, B., and Kangas, O. (2012). *Changing Social Equality: the Nordic Welfare Model in the 21st Century.* Bristol: Policy Press. doi: 10.1332/policypress/9781847426604.001.0001

Laitinen, M., Paulasto, H., and Meriläinen, L. (2020). "Samaan aikaan somessa: havaintoja englannin kielen käytöstä itäsuomalaisessa tviittivirrassa," [Meanwhile on social media: Observations of English language use in tweets in Eastern Finland]. in *Ääniä idästä: Näkökulmia Itä-Suomen monikielisyyteen* [Sounds from the East: Perspectives on the Multilingualism of Eastern Finland], eds L. Kolehmainen, H. Riionheimo, and M. Uusitupa (Helsinki: SKS).

Lane, P. (2016). Standardising Kven: participation and the role of users. *Sociolinguistica* 30:7. doi: 10.1515/soci-2016-0007

Latomaa, S. (1998). English in contact with "the most difficult language in the world": the linguistic situation of Americans living in Finland. *Int. J. Sociol. Lang.* 133, 51–72. doi: 10.1515/ijsl.1998.133.51

Latomaa, S., and Nuolijärvi, P. (2002). The language situation in Finland. *Curr. Issues Lang. Plann.* 3, 95–202. doi: 10.1080/14664200208668040

Lemberg, D. (2018). "The universal language of the future": decolonization, development, and the American Embrace of Global English, 1945–1965. *Modern Intellect. History* 15, 561–592. doi: 10.1017/S1479244317000166

Leppänen, S., and Nikula, T. (2008). "Johdanto [Introduction]," in *Kolmas Kotimainen: Lähikuva englannin käytöstä Suomessa [Third domestic language: Close-ups of the use of English in Finland]*, eds S. Leppanen, T. Nikula, and L. Kääntä (Helsinki: Suomalaisen Kirjallisuuden Seura).

Leppänen, S., and Pahta, P. (2012). "Finnish culture and language endangers – language ideological debates on English in the Finnish press from 1995 to 2007," in *Dangerous Multilingualism: Northern Perspectives on Order, Purity and Normality*, eds J. Blommaert, S. Leppänen, P. Pahta, and T. Räisänen, T (Palgrave Macmillan). doi: 10.1057/9781137283566_7

Leppänen, S., Pitkänen-Huhta, A., Nikula, T., Kytölä, S., Törmäkangas, T., Nissinen, K., et al. (2011). *National Survey on the English language in Finland: Uses, meanings and attitudes Studies in Variation, Contacts and Change in English*. Helsinki: Research Unit for Variation, Contacts and Change in English.

Levisen, C. (2012). "Are danes truly the happiest people on earth?" in *Happiness and "Pain" Across Languages and Cultures*, eds C. Goddard and Z. Ye (Amsterdam: John Benjamins).

Levisen, C. (2014). The story of "Danish Happiness": Global discourse and local semantics. *Int. J. Lang. Cult.* 1, 174–193. doi: 10.1075/ijolc.1.2.03lev

Lippi-Green, R. (2012). *English With an Accent: Language, Ideology and Discrimination in the United States (2nd ed)*. London: Routledge. doi: 10.4324/9780203348802

Loftsdóttir, K., and Jensen, L. (2012). *Whiteness and Postcolonialism in the Nordic Region: Exceptionalism, Migrant Others and National Identities*. Farnham: Ashgate.

McRae, K. D. (1997). *Conflict and Compromise in Multilingual Societies: Finland.* (Vol. 3). Wilfrid Laurier University Press.

Meriläinen, L. (2021). "The interplay between universal processes and cross-linguistic influence in the light of learner corpus data: examining shared features of non-native englishes," in *Learner Corpus Research Meets Second Language Acquisition*, eds B. Le Bruyn and M. Paquot (Cambridge: Cambridge University Press). doi: 10.1017/9781108674577.005

Norrby, C. (2015). "English in scandinavia: monster or mate? Sweden as a case study," in *Challenging the Monolingual Mindset*, eds J. Hajek and Y. Slaughter (Clevedon: Multilingual Matters). doi: 10.21832/9781783092529-004

Nuolijärvi, P., and Vaattovaara, J. (2011). "De-standardisation in progress in Finnish society?" in *Standard Languages and Language Standards in a Changing Europe*, eds K. Tore and C. Nikolas (Novus Forlag).

Oksaharju, J. (2021). *English as a Corporate Language: Employees' Perceptions of Business English Lingua Franca*. MA thesis, University of Jyväskylä.

Organisation for Economic Co-operation and Development (OECD). (2021). *Taxing Wages 2021: Finland*. Available online at https://www.oecd.org/tax/tax-policy/taxing-wages-finland.pdf

Pahta, P. (2008). "The history of english studies in Finland," in *European English Studies: Contributions towards the History of a Discipline Vol. II,* eds B. Engler and R. Haas (Leicester: The English Association for ESSE).

Pease, D. E. (2009). *The New American Exceptionalism*. Minnesota: University of Minnesota Press.

Pennycook, A. (2017). *The Cultural Politics of English as an International Language*. Routledge. doi: 10.4324/9781315225593

Peterson, E. (2019). *Making Sense of "Bad English": an Introduction to Language Attitudes and Ideologies*. London: Routledge. doi: 10.4324/9780429328343

Peterson, E. (2020). The US is not enough: Why the real world of linguistics needs your voice (Response to Charity Hudley et al.). *Language*, 96, e247–e253. doi: 10.1353/lan.2020.0076

Peterson, E. (2022). The English language in Finland: tool of modernity or tool of coloniality?

Peterson, E., and Hall, M. (2017). *English Prescriptivism in Higher Education Settings: Focus on Nordic Countries*. in *The Routledge Handbook of Linguistic Prescriptivism*, eds J. Beal, M. Lukač, and R. Straaijer (Routledge).

Peterson, E., and Vaattovaara, J. (2014). *Kiitos* and *pliis*: The relationship of native and borrowed politeness markers in Finnish. *J. Politeness Res.* 10:11. doi: 10.1515/pr-2014-0011

Piller, I. (2016). *Linguistic Diversity and Social Justice: An Introduction to Applied Sociolinguistics* (First edition). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199937240.001.0001

Pitkänen-Huhta, A., and Hujo, M. (2012). *Experiencing Multilingualism — the Elderly Becoming Marginalized? Dangerous Multilingualism: Northern Perspectives on Order, Purity and Normality*. Basingstoke: Palgrave Macmillan. doi: 10.1057/9781137283566_12

Pratt, J., and Eriksson, A. (2013). *Contrasts in Punishment: An Explanation of Anglophone Excess and Nordic Exceptionalism*. London: Routledge. doi: 10.4324/9780203096116

Ranta, K., and Kanninen, J. (2019). *Vastatuuleen. Saamen kansan pakkosuomalaistamisesta*. Helsinki: SchildtsandSöderströms.

Røyneland, U. (2009). Dialects in Norway: Catching up with the rest of Europe? *Int. J. Sociol. Lang.* 2009, 196–197. doi: 10.1515/IJSL.2009.015

Saarinen, T. (2020). *Higher Education, Language and New Nationalism in Finland: Recycled Histories* (1st ed. 2020). Palgrave Macmillan. doi: 10.1007/978-3-030-60902-3

Saarinen, T., and Taalas, P. (2017). Nordic language policies for higher education and their multi-layered motivations. *Higher Educ.* 73, 597–612. doi: 10.1007/s10734-015-9981-8

Saraceni, M. (2019). "The politics of world englishes," in *The Cambridge Handbook of World Englishes,* eds D. Schreier, M. Hundt, and E. W. Schneider (Cambridge: Cambridge University Press). doi: 10.1017/9781108349406.027

Savolainen, J. (2021). *The Grim Secret of Nordic Happiness*. Available online at https://slate.com/news-and-politics/2021/04/finland-happiness-lagom-hygge.html

Seargeant, P. (2012). *Exploring World Englishes: Language in a Global Context*. Routledge.

Sharma, D., Levon, E., Watt, D., Ye, Y., and Cardoso, A. (2019). Methods for the study of accent bias and access to elite professions. *J. Lang. Discrim.* 3, 150–172. doi: 10.1558/jld.39979

Sundberg, G. (2013). Language policy and multilingual identity in Sweden through the lens of generation Y. *Scand. Stud.* 85, 205–2332. doi: 10.5406/scanstud.85.2.0205

The Guardian (2018). *Why Nordic Countries Might Not Be as Happy as You Think*. Available online at https://www.bbc.com/news/world-europe-45308016

Vaarala, H., Riuttanen, S., Kyckling, E., and Karppinen, S. (2017). *Language reserve. Now! Follow-up on Pyykkö's Report Multilingualism into a strength*. University of Jyväskylä.

Vaattovaara, J., and Peterson, E. (2019). Same old paska or new shit? On the stylistic boundaries and social meaning potentials of swearing loanwords in Finnish. *Ampersand* 6:100057. doi: 10.1016/j.amper.2019.100057

Vik, H. H., Jensen, S. L., Lindkvist, L., and Strang, J. (2018). Histories of Human Rights in the Nordic Countries. *Nordic Journal of Human Rights*, 36, 189–201. doi: 10.1080/18918131.2018.1522750

Young, T. J., and Walsh, S. (2010). Which English? Whose English? An investigation of 'non-native' teachers' beliefs about target varieties. *Lang. Cult. Curriculum* 23, 123–137. doi: 10.1080/07908311003797627

English Features in Saudi Arabia: A Syntactic Study Within the World Englishes Framework

Nuha AlShurfa [1], Norah Alotaibi [2], Maather Alrawi [1] and Tariq Elyas [1]*

[1] Department of European Languages and Literature, King Abdulaziz University, Jeddah, Saudi Arabia, [2] English Language Department, Taif University, Taif, Saudi Arabia

Saudi English has recently emerged as a new variety within the World Englishes framework. Many scholars have argued that there is still a gap in the literature and more studies on Saudi English are needed. This study hopes to contribute to the growing research interest in Saudi English studies. The current study aims to identify Saudi English syntactic characteristics classified in relation to noun phrase, verb phrase, prepositional phrase, and clausal structure. Data were collected using three methods: conversation in natural settings, open-ended questions, and students' writings. The findings confirm that substrate-superstrate interaction affects many syntactic characteristics of Saudi English.

Keywords: Saudi English, syntactic features, World Englishes, language in contact, English in Saudi Arabia

INTRODUCTION

Saudi English (SE) is an emerging variety of English (Al-Rawi, 2012; Mahboob and Elyas, 2014; Fallatah, 2017; Elyas et al., 2021) that has not been well-researched and mostly investigated from a pedagogical perspective, where grammatical errors in English have been the focus rather than investigating them as features. However, few recent studies on SE were published within the World Englishes framework (Al-Rawi, 2012; Mahboob, 2013; Mahboob and Elyas, 2014; Fallatah, 2017; Bukhari, 2019; Elyas and Mahboob, 2021a; Elyas et al., 2021), but there is still a gap in the literature and more studies on SE are needed. This study hopes to contribute to SE literature by identifying its syntactic characteristics. To achieve this, several examples have been extracted from SE data collected in different contexts. Grammatical characteristics are analyzed based on Mesthrie and Bhatt (2008) and Kortmann and Szmrecsanyi (2009) classifications of New Englishes features.

Al-Shabbi (1989) stated that English in Saudi Arabia was first introduced in schools in 1924. In the beginning, English was taught for pupils in elementary school, but then it was restricted to intermediate and secondary stages in 1943. After that, English was taught as a subject in Islamic Law in the city of Makkah in 1949. The first English department in Saudi Arabia was established at King Saud University in 1957 [refer to Al-Haq and Smadi (1996) and Elyas (2011)]. The importance of English in Saudi Arabia increased in the 1970s in different domains, such as business, education, and government sectors. Saudi companies, such as Aramco, require their employees to be proficient in English, and different governmental institutions offer English training courses to their employees (Elyas and Picard, 2010, 2013; Elyas, 2011; Mahboob, 2013).

According to a study by Fallatah (2017), the status of English has changed since the King Abdullah scholarship program started in 2005. Many Saudi students traveled to English-speaking countries to earn higher academic degrees in many fields. The importance of English is also evident

in its use as a lingua franca between Saudis and foreigners, such as pilgrims, tourists, and workers in international companies (Elyas, 2011). Al-Rawi (2012) stated that learning English helps university graduates to increase their chances of employment in many Saudi and international companies. This, in turn, has motivated well-off families with strong socio-economic backgrounds to send their children to schools where English is taught as a second language. As such, education and proficiency in English have become part of their social prestige (Al-Rawi, 2012). As a result, the Saudi government has made a step forward to increase the number of English institutions, and English-online courses have become a trend by many Saudis who pursue better employment (Elyas and Picard, 2010, 2012, 2018; Elyas, 2011).

SAUDI ENGLISH

Recent studies have discussed the grammatical errors in English by Saudis (Elyas, 2008, 2011; Mahboob, 2013; Alahmadi, 2014; Mahboob and Elyas, 2014; Osman, 2015; Khatter, 2019; Bukhari, 2021). Their approaches lie outside the scope of this study although many of the syntactic features that they tackle overlap with those found in our data in SE. According to Mesthrie and Bhatt (2008), the investigation of English varieties is not a modern approach as it began in the nineteenth century. They state that most English varieties were studied in isolation until the 1980s. Then, many pioneering scholars called these varieties of English "World Englishes". Kachru is one of the important scholars in the field who received the main credit for founding this field of study. In a similar vein, Onysko (2022) state that "a range of studies have emerged along related strands of research concerned with the global spread and creation of Englishes (World Englishes)," where the spread of English, globalization, and explicit contact impact the "other languages and the influences that emerge from this contact" (p.1). A recent study by Bolton (2018) calls upon disciplinary debates and future directions in World Englishes. Globalization and language worlds have an immediate impact on World Englishes worldwide and contact with local languages (Bolton, 2013, 2019; Onysko, 2016). English is seen as a highly diversified language that appears in a multitude of different varieties and dialects across the globe (Siemund, 2013; Siemund et al., 2013, 2021; Vaicekauskien, 2020). In the Saudi context, we can find that the most directly related studies to Saudi English and World Englishes paradigm are those by Al-Rawi (2012), Mahboob (2013), Mahboob and Elyas (2014), Bukhari (2021), Elyas et al. (2021). In addition, there has been a growing scholarship in World Englishes in Arabian Gulf in recent years. The status of English in the Arabian Gulf was investigated and analyzed by many researchers such as Elyas and Mahboob (2021a), Hillman et al. (2021), Hopkyns et al. (2021), Siemund et al. (2021), Tuzlukova and Mehta (2021), and van den Hoven and Carroll (2021). A recent bibliography of World Englishes in the Middle East and North Africa (MENA) has been documented by Elyas and Mahboob (2021b) stressing the

importance of World Englishes by current research interests in the region.

Kachru (1985) succinctly proposed a model for World Englishes which consists of three circles of varieties. First, there are the Inner Circle varieties which involve native speakers of English as in the UK, the US, Australia, Canada, and New Zealand. Second, we find the Outer Circle varieties which involve second language Englishes such as those in Nigeria, Kenya, and India. Third, the Expanding Circle varieties refer to English as a Foreign Language (EFL) as in China, Egypt, and Saudi Arabia.

SE STUDIES WITHIN THE WORLD ENGLISHES FRAMEWORK

Al-Rawi (2012) study investigated four syntactic properties of SE following the list of features found in Kortmann and Szmrecsanyi (2004), which are also relevant to the current study. The four syntactic features are, first, *be* deletion as in "They … not able to hear anyone". The second feature is the insertion of the definite article *the* as in: "I prefer **the** fish". The third feature is the omission of the indefinite article as in "When I grew older, I want to be doctor". The fourth feature is the lack of subject-verb agreement as in "My father always **teach** me how to discover my capabilities".

Mahboob (2013) and Mahboob and Elyas (2014) also studied SE and explored Middle East Englishes with a focus on SE. They studied the features of SE by examining high school textbooks and assumed that SE was regarded as a local variety having its own features. Al-Shurafa (2014) are related in the sense that the syntax of Arabicized English and its status in the Gulf are discussed at a wider level. Fallatah (2017) conducted a study that investigated bilingual creativity in SE. Elyas et al. (2021) studied SE and translanguaging in comedy clubs. Their findings identified several features of Saudi English bilingual creativity that can fall into certain categories and recurrent themes, for example, code-switching, syntactic and semantic creativity, translation, and lexical creativity. Elyas et al. (2021) argue that Saudi nationals have been using "simple English" with a local flavor of their "Saudi English" in their day-to-day communication as part of their translanguaging and mixing English with distinct Arabic linguistic repertoire creatively.

According to Kachru (1985), bilingual creativity is a linguistic process that results from the speakers' knowledge of two or more languages. In his study, he identified six forms of bilingual creativity reflected in the language of five comedians: code switching, syntactic variation, cultural reference, pronunciation shift, lexical variation, and semantic variation.

METHODOLOGY AND THEORETICAL FRAMEWORK

The methods used to collect data for this study are described in this section. Relevant demographic information about the participants and the number of the extracted expressions from the data for scrutiny are discussed.

Participants and Data

The data were collected from 139 educated male and female Saudis who speak English as a foreign language. Most participants learned English in school, starting in the intermediate stage when they were first exposed to English. Their age is between 18 and 44. There are 114 women and 25 men of different educational levels. One participant is a Ph.D. candidate, 15 participants have a Master's degree, 69 participants are in MA and B.Sc. programmes, 48 participants are from the preparatory year, and 6 are high school students who have not yet attended college. The participants have different academic degrees, including Medicine, Engineering, Nursery, Biology, Computer Science, Law, Business Administration, Kindergarten in Arabic and English. It is worth mentioning that 32 of the participants are majoring in English.

Data Collection

The data were obtained through three methods: spontaneous language, open-ended questions, and students' writings. First, language data were obtained from a chatting conversation among five friends in a WhatsApp group. They were discussing different topics of mutual interest. The participants were undergraduates majoring in English. Second, open-ended *wh* questions were given to participants, distributed in two ways. Four questions related to the participants' professions were sent to them *via* WhatsApp. The investigator sent these questions to 41 relatives and friends. Most of the participants answered orally by sending recordings of their speech while some of them sent written answers. Twenty-seven answers were recorded and transcribed by the researchers.

Open-ended *wh* questions were sent out *via* a questionnaire which was designed online using Google Forms. It consisted of several demographic questions and five questions relating to the participants' interests. The questionnaire was distributed through Facebook to Saudi academic groups, sent *via* WhatsApp, and shared with the staff at the Security Forces Hospital in Makkah. The number of participants obtained through this method was 54. However, only 49 of the respondents were used; four participants were excluded because they were not Saudi nationals, and one was excluded because her answers were written in Arabic. Twenty-nine sentences were extracted for this analysis. The problem with this method is that some participants gave short answers that only contained one or two words and some left questions unanswered.

The last method involved the collection of students' writings. The articles of 90 students studying in the preparatory year at college were collected. All of them were women, ranging in age between 18 and 21 years and studying various scientific courses besides the English language. They study English for 18 h a week and are either beginners or intermediate-level speakers. Beginners were asked by their English instructors to write short paragraphs about themselves to be read out in class, and they were collected immediately. Advanced-level students wrote about their favorite people and special celebrations. In total, 180 sentences were obtained for analysis. This method was effective because students wrote longer sentences which helped to elicit various features.

Data Analysis

The data were analyzed based on the structural categories of World Englishes that were presented in Mesthrie and Bhatt (2008). They divided the features found in World Englishes into two categories. The first category relates to the lexical and phrasal levels within the sentence while the second pertains to the clause level. The first category is divided into five subsections: noun phrase, verb phrase, prepositions, conjunctions, and *wh*-words. The second category is subdivided into nine sections within the clause level: word order, relative clauses, passive, comparison, tag questions, answers to yes/no questions, adverb placement, and other constructions that are limited to a few varieties (Mesthrie and Bhatt, 2008). In addition to Mestherie and Bhatt's classification, the data of this study was analyzed based on Kortmann and Szmrecsanyi (2004) scheme of categorization. On their list, they provided 76 features found in non-standard Englishes around the world. All of these features were listed under 11 major categories: pronouns, noun phrases, verb phrases, adverbs, negation, agreement, relativization, complementation, discourse organization, and word order (Kortmann and Szmrecsanyi, 2004, p. 1,146–1,148). The current study attempts to identify the syntactic characteristics of SE that were found in the collected data and examines those against the lists of the previous categories.

RESULTS AND DISCUSSION

The syntactic characteristics found in the variety of SE are presented and discussed below. They involve divisions and subdivisions of noun phrase structures, verb phrase structures, prepositions, and clause structures. Related examples elicited by the participants are also provided.

The Noun Phrase
Articles

Most participants demonstrated an irregular use of articles. For example, they tended to omit the indefinite articles *a* and *an*. This is a common feature of World Englishes. As illustrated in examples (1), (2), (3), and (4) below, the speakers omitted the indefinite articles:

(1) It is … international language.
(2) I was… outstanding student.
(3) I am… talented speaker.
(4) It is … hard question.

Interestingly, the speakers who uttered sentences (1) and (2) were majoring in English: the first speaker was a student pursuing a BA in English, and the second was a postgraduate student. The speakers of the last two sentences, (3) and (4), were majoring in other disciplines. The results of the first two methods, namely: natural speech and open-ended questions, were compared to the results from the third method (students' writings) because most participants involved in the first two methods were more educated than those involved in the third method. The students who participated in the third method were in their preparatory year of learning English. More examples are provided below where the indefinite article *a/an* are omitted:

(5) I am… student in the university KAU.
(6) She was… English student.
(7) She was …job assistant.
(8) I don't have… job.

Examples (5)–(8) exhibit the omission of *a/an* before the singular count nouns *student*, *English student*, *job assistant*, and *job*, respectively.

Moreover, the data showed that SE speakers used the article *a* before plural nouns, as examples (9) and (10) demonstrate.

(9) Each one have *a* different roles.
(10) I love watching *a* movies.

It is also observed that SE speakers tended to add indefinite articles even if the noun is preceded by an adjective modifier as in (9) above.

Another feature that was found in SE pertains to the redundant insertion of the definite article *the*, as illustrated in (11) – (15) below:

(11) I also like *the* novels.
(12) I believe that *the* women have their rights.
(13) I like cook and read *the* books.
(14) It is *the* very boring job.
(15) I like how you simplify *the* informations.

In examples (11), (12), and (13), the speakers inserted *the* before nouns, which is considered a deviation from Standard English. In (14), the definite article *the* was used irregularly, preceding the adverb *very* and defining a general noun "*job*". The last example (15) was used by a participant of the third group who used the definite article *the* with the abstract noun *information(s)*. This variation in the use of the definite article was also found in the syntax of many World Englishes (Mesthrie and Bhatt, 2008).

According to Al-Rawi (2012), SE speakers demonstrate this type of variation in their speech as a result of their native substrate effect. The system of the Arabic language does not involve the use of indefinite articles. Definiteness and the introduction of a topic that is already known to both interlocutors are expressed by the use of *?al-* "the".

Number

Participants also showed variation in their use of plural forms. Several SE speakers did not add the plural marker *-s* to pluralize nouns, as exemplified in (16) and (17). However, it is clear from the context of these sentences that plurality should be expressed, i.e., *books and interests*.

(16) I love reading *book*.
(17) Swimming, jogging, and cooking are my *interest*.

Moreover, a singular form of the noun is used instead of the plural form. In examples (18) and (19), the numerals *one of* and *four* before the nouns indicate that the respective nouns are plural. However, the plural marker *-s* was not used on the following noun as shown by the participants of the second group.

(18) I have four *sister*.
(19) One of my *sister* is married.

In (20)–(22), although participants used the quantifiers *few*, *a lot*, and *many*, the nouns that followed these quantifiers were expressed in the singular, as they were not marked with *-s*.

(20) There's few *thing* I'd like to change.
(21) A lot of *thing*.
(22) I work in many *city*.

In addition to plural *-s* omission, SE speakers may add the *-s* to singular nouns. According to Mesthrie and Bhatt (2008) this is called an overgeneralization of rules, as in (23).

(23) Each one have a different *roles*.

In (24), the plural form of the noun *woman,* which is *women*, is irregular, but the speaker did not change the noun and used the unmarked noun *woman*. The use of unmarked or singular nouns instead of the irregular form of the plural noun occurs a few times in the current data.

(24) *Woman* become as men.

In examples (25) and (26) below, the participants used the non-count nouns *information* and *communication*. In Standard English, these two nouns are uncountable and do not take plural *-s*. However, the data showed that SE speakers may add the plural marker *-s* to uncountable abstract nouns.

(25) I like how you simplify the *informations*.
(26) We have strong social *communications*.

The substrate system (L1) of the speakers' native language affected the use of the plural form by SE speakers. For example, the equivalent of the abstract nouns *information* and *communication* are considered countable nouns in Arabic. As a result, SE speakers tend to add plural markers to these nouns.

Possession

Within the current data, possessive *'s* omission was not common. Only one example was found in (27) below:

(27) I want to change people… mentality.

In (27), the possessive marker on to the noun *people is omitted*. Interestingly, this sentence was produced by a female who majored in English and had knowledge of the language. This can be evidence that this feature requires more investigation in future studies. According to Mesthrie and Bhatt (2008), the possessive feature in New Englishes did not receive much attention. Platt et al. (1984) found that there is a deletion of *'s* when New Englishes speakers express possession (as cited in Mesthrie and Bhatt, 2008), which is also common in African-American English.

Examples (28)–(33), below, indicate various features of possessive pronoun substitution, which is different from Standard English.

(28) My sister *her* name Waffa.
(29) My best friend *she* name is Amal.
(30) *She* name my mom is Souad.
(31) *She* name is Naem.
(32) *She's* name Aseel.

(33) My mother *is* name Misaa.

These substitutions are used many times by different speakers in the second group. In (28), the speaker substitutes the object pronominal *her* for possessive's, so *my sister's name* becomes *my sister her name*. In (29), the speaker uses the subject pronoun *she* instead of 's. In both (30) and (31), the speakers use *she* to express possession but with different word orders. Interestingly, adding 's to the pronoun *she* instead of adding it to the head noun. According to Lightbown and Spada (2013, p.47) "the possessive's is one of the grammatical morphemes that is not acquired early in the process of learning English".

Pronouns

Variations in the use of pronouns are also found in SE as illustrated (34–42) below:

(34) My plan for the future to make *me* be better.
(35) I development *me* about everything.
(36) I know *she* in school.
(37) Correct *me* wrong thoughts.
(38) It was hard for *my* when I come to Jeddah.
(39) *She* name my mom is Souad.
(40) *He* name is brother Talal.
(41) *Me* and my cousin playing.
(42) Aseel *it* is my best friend.

In (34) and (35) above, the object pronoun *me* is used instead of *myself*. In (36), the pronoun *she* is used instead of the object pronoun *her*. SE speakers may use object pronouns in place of possessive pronouns. In (37) above, the participant uses the object pronoun *me* before the noun where the possessive pronoun *my* is expected to occur. In (38) above, the speaker uses the possessive pronoun *my* instead of *me*. In (39) and (40) above, the subject pronouns *she* and *he* are used instead of *her* and *his*, respectively. In (41) above, the object pronoun *me* is used instead of the subject pronoun *I*. Finally, in (42) above, *it* is inserted.

Table 1 below summarizes the SE characteristics in the nominal domain:

TABLE 1 | Saudi English (SE) characteristics in noun phrases.

1	Omission of indefinite article:	*I was outstanding student.*
2	Insertion of definite article:	*I like **the** novels.*
3	Insertion of indefinite article:	*Each one have **a** different roles.*
4	Use of definite article with non-count nouns:	*I like when you simplify **the** informations.*
5	Deletion of -s plural marker:	*I work in many **city**.*
6	Insertion of plural marker with singular nouns:	*Each one have a different **roles**.*
7	Insertion of plural -s with non-count nouns:	*I like how you simplify the **informations**.*
8	Deletion of possessive's marker:	*I want to change people mentality.*
9	Use of object pronouns for reflexive:	*I developed **me** about everything.*
10	Use of *me* in place of *I*:	***Me** and my cousin playing.*
11	Irregular use of possessive pronouns:	*It was hard for **my** when I came to Jeddah.*

The Verb Phrase
Tense
Saudi English (SE) speakers, furthermore, showed variation in the use of different tenses. One of these features is the use of the unmarked verb in the past tense. This feature was frequently used in the speech of SE speakers. All the italicized verbs in the following examples are uninflected although they refer to temporal periods located before the moment of speaking.

(43) I just *wake* up.
(44) After finishing high school, I *decide* to major in English.
(45) I *visit* a lot of places.
(46) I was prepared and *wear* my simple make up and I *do* my hair.
(47) I *send* you message 2 weeks ago.

In (43), the speaker substituted *wake up* for the irregular past *woke up*. In (44) above, the speaker refers to a decision that was made in the past, as the context shows, and deleted the past marker -ed from the verb *decide*. In (45), the speaker talks about a trip that occurred in the past, but the verb is equally not marked with -ed. In (46) above, the speaker began the sentence in the past tense; then, the speaker used unmarked verbs in the following sentences. The speaker only used the -ed marker with the first regular verb, but did not use the irregular forms of *wear* and *do* and kept them unmarked. In (47), the speaker specified the time of the action as 2 weeks *ago*, which indicated that it happened in the past. However, the speaker used the unmarked verb *send* instead of the irregular form *sent*. This feature was used by SE speakers regardless of their educational levels.

Another variable use of tense marking occurs in reference to future time. According to Mesthrie and Bhatt (2008), speakers of World Englishes may use present and future tenses interchangeably. The data of the current study shows that there is no overlapping between present and future in SE. The following examples contain "*will*" to express the future tense, but the structure of their utterances was different from Standard English as in (48–51).

(48) I will *be* improve my English language.
(49) I will *be* finish from exams.
(50) I will *to* be hero nurse.
(51) The answer … be long.

In (48) and (49), the speakers added *be* to *will* to express the future tense. They added it to the sentence even though there is a main verb. In (50), the speaker inserted a redundant preposition. In (51), *will* was omitted by the speaker although she intended to use future tense.

Aspect
The perfect aspect in Standard English (*have* + past participle verb) is attested in the data. However, some World Englishes show variation in perfect aspect use. SE speakers, for instance, were also found to exhibit variation in aspectual usage, as in (52)–(55).

(52) I *know* her five years ago by accident.

(53) I recently *moved* to Jeddah.
(54) She never *lie* to anyone.
(55) Have you ever *meet* her?
(56) I *don't* fix my laptop yet.

In (52), (53), (54), (55), and (56) above, the speakers refer to aspectual events that started in the past and are still happening in the present by using the presents tense. In these contexts, the present perfect (*have/has* + past participle verb) is expected to be used instead of only the present or the past.

Modality

Modal verbs are used in English for the expression of the speaker's perspective. For example, it can say something about the "speaker's attitude toward the action that took place" [(Mesthrie and Bhatt, 2008), p. 64]. SE speakers showed variation in the use of modals. First, *could* is used in Standard English for the past of *can* which is used to express ability. Also, *could* is used to ask for permission. The participants of this study were found to use *could* instead of *can*. Variation in the expression of modality in SE is exhibited in (57)–(60).

(57) I *could* learn.
(58) I *could* speak a bit Turkish.
(59) We have strong social communications which *could* be good.
(60) They *have should* do a lot.

In (57), *could* is used to express her ability to learn something in the future. In (58), the speaker told the listener about her ability to speak Turkish. Both sentences are not in the past. However, the speakers used *could* to express their ability. SE speakers also used *can* instead of *will* to express future tense as (59) illustrates. In Standard English, *may* is used for probability, but it may be replaced with *could* in SE speech as in (59). The speakers in (60) insert the aspectual *have* next to the modal *should*.

Agreement

Lack of subject-verb agreement in the present tense is one of the most common features in many varieties of New Englishes (Mesthrie and Bhatt, 2008). It was also found as a common feature of SE based on the collected data of the current study.

(61) It *give* us insight for the truth.
(62) He always *take* me to the mall.
(63) She *know* how to cook.
(64) As my dad always *say*.
(65) He *have* a nice smile.
(66) She *do* not like animals.
(67) I *knows* her about 3 years.
(68) Dedication and honesty *plays* a big part.

Examples (61), (62), (63), (64), (65), and (66) show that the verb lacks the third person singular inflection to agree with the singular subject while in (67) and (68), the speakers add the inflection –*s* where it is not required. Al-Rawi

(2012) points out that the non-standard agreement is highly frequent among SE speakers regardless of their level of education.

Forms of the Verb Be

Omission of the verb *be* is considered one of the commonly occurring characteristics of World Englishes in SE (Elyas, 2011; Al-Rawi, 2012; Mahboob, 2013; Mahboob and Elyas, 2014; Fallatah, 2017; Elyas et al., 2021). The current data also showed that participants delete copular *be* in the present tense as in the following examples:

(69) He… 10 years old.
(70) I… from Saudi Arabia
(71) We… good friends.
(72) I …… in the library.
(73) They …. twins.
(74) She ….. happy.

In (70)–(74), participants deleted the copular verb *be* that connects subjects with copulative complements.

Saudi English speakers further tend to drop the auxiliary *be*. The auxiliary *be* is an aspectual head that selects a progressive verb inflected with –*ing* as shown below.

(75) My father… working in King Abdulaziz University.
(76) She… doing her mother work.
(77) She… coming to me home.

Examples (75)–(77) show that participants drop the auxiliary form *is* before the progressive verbs *working, doing,* and *coming*.

Another form of auxiliary *be*, namely, the one preceding the passive verb is also deleted in SE as illustrated in (78) and (79).

(78) It should … done by all of us.
(79) He… saved by some of his teachers.

The sentences in (78) and (79) both lack the auxiliary be before the passive verbs *done* and *saved*.

Saudi English participants were also found to insert the verb *be* before the verbs in several sentences as manifested in the sentences below.

(80) It *is* seem that we cannot do it.
(81) He *is* read books every day.
(82) I *am* like the idea.
(83) We *are* totally agree with you.

Participants insert *be* where it is not expected to be used. In (80) and (81), *is* inserted. In (82), *am* is inserted. In (83), *are* is inserted.

Table 2 below summarizes the verb phrase characteristics in SE.

Prepositions

Saudi English speakers also showed variation in the use of prepositions in the current study.

(84) I went …. Makkah.
(85) I argue ….. the idea.
(86) I really enjoyed *in* the movie.

TABLE 2 | Saudi English characteristics in verb phrases.

1	Irregular use of past tense:	*I send you message two weeks ago.*
2	Irregular use of future:	*The answer be long.*
3	Irregular use of perfect aspect:	*I recently moved to Jeddah.*
4	Irregular use of modals:	*I could speak a bit of Turkish.*
5	Lack of subject-verb agreement:	*She know how to cook.*
6	Deletion of Be:	*I from Saudi Arabia.*
7	Be insertion with the main verb:	*I am like the idea.*

(87) He admitted *by* his mistake.
(88) I looked *from* the window.

Three types of variation are evident in (84)–(88). One type is the zero-use of the preposition. In (84) and (85), the preposition *to* is dropped, and in (86), the preposition *for* is dropped. Another type is the insertion of the preposition. In (86) and (87), *in* and *by* are inserted. The third type is the replacement of a preposition by another. In (88), the preposition *through* is being replaced by *from*. **Table 3** summarizes SE characteristics of preposition use, which may indicate possible L1 influence as found in the data.

The Structure of SE Clauses
Omission of Subjects
Some SE speakers were found to omit the subject in the sentence, as in (89) and (90).

(89) Defend my country.
(90) She live in Makkah, but … come to Jeddah every day.

In (89), the speaker omitted the subject before the verb *defend*. In (90), the clause is left with a null subject before the verb *come*.

Use of Repeated Subjects
Participants repeat the subject by redundantly inserting a pronoun immediately after the subject.

(91) Men and women *they* have roles to play.
(92) All the women *they* drive.

In (91), although the speaker used coordinate subjects, the pronoun *they* is inserted after the subject forming a repetition of the subject. In (92), the subject (*all the women*) is followed by a redundant pronoun *they* that co-refers with the subject. This is a clear influence of the underlying L1 as in modern Arabic a subject pronoun is added after the subject to indicate an emphasis on the subject being discussed.

Omission of Objects
Saudi English speakers were also found to omit the object pronoun with transitive verbs. In (93), the participant left the place of the object empty as follows:

(93) I like … to put in their mind.

TABLE 3 | Saudi English characteristics of prepositions.

1	Preposition deletion:	*We traveled … Dubai.*
2	Preposition insertion:	*I really enjoyed in the movie.*
3	Preposition substitution:	*I looked from the window.*

TABLE 4 | Structure of SE clauses.

1	Omission of subjects	*Defend my country.* [as a statement]
2	Use of repeated subjects	*All women **they** drive.*
3	Omission of objects	*I like to put in their mind.*
4	Irregular use of superlative	*She is **biggest** than me.*
5	Irregular use of comparative	*My father is a **best** man in the world.*

Comparative and Superlative Interchanging
The data also show variation in the use of comparative and superlative structures, as illustrated by the following examples.

(94) She is *biggest* than me.
(95) My father is *a* best man in the world.

First, the speaker in (94) applied the superlative in a comparative construction, as she added the morphological form *-est* to the adjective *big*. Second, the speaker in (95) used the indefinite article *a* instead of the definite article *the* when applying the superlative structure. However, these features were not frequently found in the collected data. As a result, further investigation is needed. **Table 4** below summarizes the four characteristics found in our data in SE clauses.

CONCLUSION AND RECOMMENDATION

To sum up, this study was conducted to identify the syntactic characteristics of SE which is considered an outer circle variety of World Englishes. This investigation examined the speech of several Saudi speakers in order to describe SE. The results of this study show that there are several grammatical characteristics in SE speech. In relation to noun phrases, it was found that speakers show variation in the use of articles, the plural, possession, and pronouns. In verb phrases, they showed variation in the use of tense, aspect, modality, number, and forms of *be*. SE speakers were also found to use prepositions differently from Standard English speakers as they delete, insert, and substitute prepositions. In relation to the structure of English clauses, SE speakers were found to omit subjects and objects, repeat subjects, and use comparative and superlative structures irregularly. These findings confirm the results found in previous studies in SE literature (Elyas, 2011; Al-Rawi, 2012; Mahboob, 2013; Al-Shurafa, 2014; Mahboob and Elyas, 2014; Fallatah, 2017; Elyas et al., 2021). Finally, further investigation is needed to elicit more syntactic characteristics of SE. Also, other aspects of language such as phonological and lexical features of SE need to be examined.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

TE and NAlS has worked on WE and Saudi English background, history, and paradigm. NAlo has collected the data. MA has done the analysis and conclusion. All authors contributed to the article and approved the submitted version.

REFERENCES

Alahmadi, N. S. (2014). Errors analysis: A case study of Saudi learner's English grammatical speaking errors. *Arab World Engl. J.* 5, 84–98. Available online at: https://www.awej.org/images/AllIssues/Volume5/Volume5number4Decmber/6.pdf

Al-Haq, F. A. A., and Smadi, O. (1996). The status of English in the kingdom of Saudi Arabia (KSA) from 1940-1990. *Contrib. Soc. Lang.* 72, 457–484.

Al-Rawi, M. (2012). Four grammatical features of Saudi English: charting the influence of Arabic on the syntax of English in Saudi Arabia. *Engl. Today* 28, 32–38. doi: 10.1017/S0266078412000132

Al-Shabbi, A. (1989). *An Investigation Study of the Practical Preparation in EFL Teacher Preparation Programs in Colleges of Education in the Saudi Arabia* (Unpublished PhD thesis), University of Wales, Cardiff.

Al-Shurafa, N. S. (2014). On the emergence of a gulf English variety: a sociocultural approach. *Buckingham J. Language Linguistics* 7, 87–100. doi: 10.5750/bjll.v7i0.924

Bolton, K. (2013). "World Englishes, globalisation and language worlds," in *Nils-Lennart Johannesson, Gunnel Melchers and Beyza Björkman*, eds *Of Butterflies and Birds, of Dialects and Genres: Essays in Honour of Philip Shaw* (Stockholm: Acta Universitatis Stockholmiensis), 227–251.

Bolton, K. (2018). "World englishes: Disciplinary debates and future directions," in *The Routledge Handbook of English Language Studies*, 1st Edn, eds P. Seargeant, A. Hewings, and S. Pihlaja (London: Routledge), 59–76.

Bolton, K. (2019). "World Englishes: current debates and future directions," in *The Handbook of World Englishes*, 741–760. doi: 10.1002/9781119147282.ch41

Bukhari, S. (2019). *Complexity Theory Approaches to the Interrelationships Between Saudis' Perceptions of English and Their Reported Practices of English* (Doctoral dissertation, University of Southampton, UK).

Bukhari, S. (2021). English Language Teachers' Views on 'Saudi English': errors vs. variants. *Taibah University J. Arts Humanities* 28, 464–496.

Elyas, T. (2008). The attitude of American English within the Saudi Education System. *Novitas-ROYAL* 2, 28–48.

Elyas, T. (2011). *Diverging Identities: A'contextualised' Exploration of the Interplay of Competing Discourses in Two Saudi University Classrooms* (Doctoral Dissertation, University of Adelaide, Australia).

Elyas, T., AlZahrani, M., and Widodo, H. P. (2021). Translanguaging and 'Culigion' Features of Saudi English. *World Englishes* 40, 20–40. doi: 10.1111/weng.12509

Elyas, T., and Mahboob, A. (2021a). World Englishes in the Middle East and North Africa (MENA). *World Englishes* 40, 154–158. doi: 10.1111/weng.12504

Elyas, T., and Mahboob, A. (2021b). Englishes in MENA region: a contemporary bibliography. *World Englishes* 40, 290–296. doi: 10.1111/weng.12515

Elyas, T., and Picard, M. (2010). Saudi Arabian educational history: impacts on English language teaching. *Educ. Business Soc.* 3, 24–40. doi: 10.1108/17537981011047961

Elyas, T., and Picard, M. (2013). Critiquing of higher education policy in Saudi Arabia: Towards a new neoliberalism. *Educ. Business Soc.* 6, 31–41. doi: 10.1108/17537981311314709

Elyas, T., and Picard, M. (2018). "A brief history of english and english teaching in Saudi Arabia," in *English as a Foreign Language in Saudi Arabia* (New York, NY: Routledge), 70–84.

Elyas, T., and Picard, P. (2012). Teaching and Moral Tradition in Saudi Arabia: A Paradigm of Struggle or Pathway towards Globalization? *Procedia* 41, 1083–1086. doi: 10.1016/j.sbspro.2012.06.782

Fallatah, W. (2017). Bilingual creativity in Saudi stand-up comedy. *World Englishes* 36, 666–683. doi: 10.1111/weng.12239

Hillman, S., Selvi, A. F., and Yazan, B. (2021). A scoping review of world Englishes in the Middle East and North Africa. *World Englishes* 40, 159–175. doi: 10.1111/weng.12505

Hopkyns, S., Zoghbor, W., and Hassall, P. J. (2021). The use of English and linguistic hybridity among Emirati millennials. *World Englishes* 40, 176–190. doi: 10.1111/weng.12506

Kachru, B. B. (1985). "Standards, codification and sociolinguistic realism: the English language in the outer circle," in *English in the World: Teaching and Learning the Language and Literatures*, eds R. Quirk and H. G. Widdowson (Cambridge: Cambridge University Press), 11–30.

Khatter, S. (2019). An analysis of the most common essay writing errors among EFL Saudi female learners (Majmaah University). *Arab World Engl. J.* 10, 364–381. doi: 10.24093/awej/vol10no3.26

Kortmann, B., and Szmrecsanyi, B. (2004). "Global synopsis: Morphological and syntactic variation in English," in *A Handbook of Varieties of English: A Multimedia Reference Tool. Volume 1: Phonology. Volume 2: Morphology and Syntax*, eds B. Kortmann and W. E. Schneider (Berlin; Boston, MA: De Gruyter Mouton). doi: 10.1515/9783110197181

Kortmann, B., and Szmrecsanyi, B. (2009). The morphosyntax of varieties of English worldwide: a quantitative perspective. *Lingua* 119, 1643–1663. doi: 10.1016/j.lingua.2007.09.016

Lightbown, P. M., and Spada, N. (2013). *How Languages are Learned, 4th Edn.* Oxford: Oxford University Press.

Mahboob, A. (2013). Englishes of the Middle East: a focus on the Kingdom of Saudi Arabia. *Middle East Handbook of Applied Linguistics* 1, 14–27.

Mahboob, A., and Elyas, T. (2014). English in the Kingdom of Saudi Arabia. *World Englishes* 33, 128–142. doi: 10.1111/weng.12073

Mesthrie, R., and Bhatt, R. M. (2008). *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

Onysko, A. (2016). Language contact and world Englishes. *World Englishes* 35, 191–195. doi: 10.1111/weng.12190

Onysko, A; and Siemund, P. (2022). Englishes in a Globalized world: exploring contact effects on other languages. *Front. Commun.*

Osman, M. A. A. R. (2015). *Analysis of Grammatical Errors in Writings of Saudi Undergraduate Students (Doctoral dissertation)*. Sudan University of Science and Technology, Sudan.

Platt, J. T., Weber, H., and Lian, H. M. (1984). *The New Englishes*. London: Routledge.

Siemund, P. (2013). *Varieties of English: A Typological Approach*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139028240

Siemund, P., Al-Issa, A., and Leimgruber, J. R. (2021). Multilingualism and the role of English in the United Arab Emirates. *World Englishes* 40, 191-204. doi: 10.1111/weng.12507

Siemund, P., Gogolin, I., Schulz, M. E., and Davydova, J. (2013). *Multilingualism and Language Diversity in Urban Areas: Acquisition, Identities, Space, Education, Vol. 1*. Amsterdam: John Benjamins Publishing. doi: 10.1075/hsld.1

Tuzlukova, V., and Mehta, S. R. (2021). Englishes in the cityscape of Muscat. *World Englishes* 40, 231–244. doi: 10.1111/weng.12510

Vaicekauskien,e, L. (2020). The social meaning potential of the global english based on data from different communities. *Taikomoji kalbotyra* 14, 183–208. doi: 10.15388/Taikalbot.2020.14.13

van den Hoven, M., and Carroll, K. S. (2021). English-medium policy and English conversational patterns in the

UAE. *World Englishes* 40, 0205–218. doi: 10.1111/weng. 12508

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Reassembling the Pimped Ride: A Quantitative Look at the Integration of a Borrowed Expression

Stefano De Pascale[1], Dirk Pijpops[2], Freek Van de Velde[1] and Eline Zenner[1]*

[1] Quantitative Lexicology and Variational Linguistics (QLVL), Faculty of Arts, KU Leuven, Leuven, Belgium, [2] LILITH, Faculty of Philosophy and Arts, University of Liège, Liège, Belgium

Over the past decades, research on the linguistic impact of globalization has foregrounded the socio-pragmatic meaning potential and mental categorization of anglicisms, looking for signs of agentivity and contextual sensitivity in the way receptor language users incorporate borrowed English resources into their speech, both in form and in function. This brought attention to understudied phenotypes of contact-induced variation and change that go beyond the borrowing of individual lexical items (loanwords) from English. This paper aims to contribute to this endeavor, illustrating the potential of construction grammar to uncover the integration of borrowed chunks. In focus is the emergence of the verb *pimpen* "to pimp" in Dutch, a rapid innovation from the English proper name *Pimp My Ride.* A sample of 4,561 Dutch tweets containing (strings of) *pimp* posted between January 2007 and April 2020 was coded manually for formal and semantic properties. This allowed us to calculate an aggregate score of "deconstructionalization" both within and outside of the target construction [*pimp* POSS N]. Results indeed reveal a gradual blurring of the sharp contours of the construction, but also indicate that this process mainly affects the instantiations closest to the original. Linked up with the mediatized origin of the construction, our results add to our understanding of the relationship between media, language contact, and what is referred to as glocalization.

Keywords: *Pimp My Ride*, anglicisms, Dutch, construction grammar, glocalization

BACKGROUND

Globalization, broadly defined as the "intensified flows of capital, goods, people, images and discourse around the globe, driven by technological innovations mainly in the field of media and information and communication technology" (Blommaert, 2010, p. 13), has left a clear imprint on all aspects of society, including its language use. For one thing, globalization is readily linked up to the worldwide spread of English, both as *lingua franca,* the means of communication between speakers with different mother tongues in various discourse domains (e.g., Ammon, 2001 for science, Mauranen and Ranta, 2009 for business), and as a prime resource for borrowing (e.g., Onysko, 2007). In Western Europe in particular, a surge of borrowing from English was witnessed after the Second World War, resulting from the strong presence of English in mass media and pop culture (Leppänen, 2007). Initially, scholars mainly aimed to assess the impact of English influence on the lexical stock of the receptor language by tallying types and tokens of English loans in different word classes (e.g., Posthumus, 1986; Yang, 1990; Görlach, 2003), but in the past decades, there has

been a shift toward more socio-pragmatic and cognitive analyses of the linguistic manifestation and development of contact with English.

In this more recent work, researchers have emphasized the importance of usage-based analyses of the negotiation between English as the linguistic signpost of globalization and the local host languages—which we can consider as a type of "glocalization" (Androutsopoulos, 2010; Garley, 2018). The goal is to uncover (1) the characteristics of users and contexts of English insertions (e.g., Zenner et al., 2014; Vaattovaara and Peterson, 2019); (2) the semantic and socio-pragmatic nuances offered by English loans, often in comparison with alternative lexicalizations available in the receptor language (e.g., Onysko and Winter-Froemel, 2011); (3) the impact of using English insertions on the way messages and their senders are being evaluated (Van Meurs, 2010 on attitudes toward English in Dutch job ads); (4) agentivity and creativity in how different linguistic manifestations of English influence are embedded in the receptor language, both in form and in function (e.g., Andersen, 2014; Peterson and Beers Fägersten, 2018; Onysko, 2021). This paper aims to add to this latter line of research, which has opened up attention for understudied phenotypes of contact-induced variation and change that go beyond the borrowing of individual lexical items (loanwords) from English. In particular, we tie in with recent studies describing the way in which (semi-)fixed expressions and chunks of English material are integrated in the receptor language use, presenting an analysis of the local lifecycle of the globally mediatized phrase *Pimp My Ride*.

In the remainder of this background section, we lay the necessary groundwork for our study. First, we sum up the benefits of the construction grammar framework for the description of this type of contact-induced variation and change that surpasses the level of the individual word (see also Boas and Höder, 2018, 2021), drawing comparisons and uncovering differences with phraseological approaches to borrowing (Section English Phrases and Constructions in Contact). Next, we introduce the specific case study under scrutiny through a description of the results of an early pilot study on the same construction (Van de Velde and Zenner, 2010), viz. the construction *Pimp My Ride* and the derived Dutch verb *pimpen* "to pimp" (Section *Pimp My Ride*). Section Research Questions then identifies the main research questions of this paper. The data, coding procedure and analytic procedure are presented in Section Methodology, after which we describe the results in Section Results. Section Discussion and Conclusion summarizes the implications of our results for our understanding of English as a global source for contact-induced variation and change.

English Phrases and Constructions in Contact

Researchers interested in lexical borrowing have long mainly restricted their attention to isolated loanwords, perhaps as an epiphenomenon of the traditional quest for sharp dividing lines between lexical borrowing and codeswitching and the ensuing debate on the status of single word switches (Poplack, 1980; Myers-Scotton, 2002; and see Zenner and Geeraerts, 2015). After all, the most prototypical instances of lexical borrowing involve exactly such loanwords, according to Matras (2009,

p. 113): "the regular occurrence of a structurally integrated, single lexical item that is used as a default expression, often a designation for a unique referent or a grammatical marker, in a monolingual context." However, when embracing the idea that there is a continuum from borrowing to codeswitching rather than a sharp dichotomy between both (see Matras, 2009; Zenner and Geeraerts, 2015; Zenner et al., 2019), understudied contact phenomena that are part of the linguistic reality positioned between the outer poles of the prototypical codeswitch and the prototypical loanword instead come to the fore. For instance, attention has been awarded recently to the way language users adapt and integrate larger semi-fixed chunks of source language material into their own language (Andersen, 2020a).

We consider both phraseology and construction grammar[1] to be useful theoretical frameworks for this endeavor. Phraseology research aims to study the form-meaning characteristics of word combinations (Cowie, 2001; Wray, 2002; Feyaerts, 2006), viz. of all structures including minimally two words (amongst others collocations, idiomatic expressions, phrasal verbs, slogans). Construction grammar, in turn, aims to identify and compare recurring linguistic form/meaning combinations on different levels of schematicity across the lexical, phonetic and grammatical domains of language use (Fillmore, 1988; Goldberg, 1995; Croft and Cruse, 2004; Steels, 2011; Boas, 2013). Both frameworks have been applied broadly and have hence each fragmented into separate subfields, making it fairly challenging if not impossible to arrive at a fit-for-all list of necessary and sufficient criteria to define and delineate the paradigms. What is clear, however, is that although both phraseology and construction grammar have largely developed independently of each other (see Gries, 2008; Ziem, 2018), they terminologically and conceptually share a number of properties. This is for instance true when taking the subframework of Cognitive Construction Grammar (Goldberg, 1995, p. 4; Boas, 2013) as point of departure, as this paper aims to do.

Where phraseology research aims to uncover a language's *phrasicon*, viz. the inventory of phrasemes, Cognitive Construction Grammar is concerned with the *constructicon*, viz. the structured inventory of constructs or phrase types that are captured by linguists descriptively as constructions. Both constructions and phrasemes are said to be (i) fairly fixed in terms of form and meaning, though allowing for variability in some of their elements ("open slots", e.g., *Oh my* X)[2]; (ii) entrenched as units in the language users' minds; (iii) with degrees of entrenchment depending on frequency of exposure to

---

[1]Construction grammar is often written with capitals, as a way to highlight the status of construction grammar as a theory, as originally developed by Fillmore, Kay and collaborators (see e.g., Fillmore, 1988). We have chosen not to use capitalization in this paper for two reasons: (1) we do not intend to refer solely to Fillmorian Construction Grammar, but rather refer to the broad family of constructional theories, and hence only use capitals when referring to specific theories such as Cognitive Construction Grammar or Sign-based Construction Grammar; (2) we would not want to create the impression that we are promoting construction grammar over phraseology, an equally valid framework.

[2]Note that other subframeworks of construction grammar consider the occurrence of an open slot obligatory (see Michaelis, 2019). Cognitive Construction Grammar (but see also Fluid Construction Grammar; Steels, 2011, p. 3–4) define "construction" to broadly mean "symbolic unit", including words, multi-word expressions and schematic constructions with open slots (Goldberg, 2006, p. 18).

the unit in language use; (iv) semantically non-decomposable, in the sense that a complex expression can have a meaning that is not attributable to the meanings that the subparts would have independently (Kay and Michaelis, 2011) (the phraseme *kick the bucket* is not about kicking a bucket, nor does *they worked their butts off* as an instantiation of the *body part off*-construction necessarily concern the agent's behind (Goldberg, 1995; Sawada, 2000; Granger, 2009); another example is the fully schematic double object or ditransitive construction *V NP NP*, which is associated with the meaning "transfer of possession", despite it containing no lexical elements whatsoever; Goldberg, 1999, p. 199). Not surprisingly then, linguists have also relied on both paradigms to study the borrowability of semi-fixed chunks as grey zones in between prototypical lexical borrowing and prototypical code-switching. In this application, some pivotal differences between the paradigms can, however, be uncovered.

Pointing out phraseological borrowing as a largely unexplored area, Fiedler (2017) aims for an inventory of English-based phrasemes in German, classified according to formal parameters of the borrowed phrasemes, mainly contrasting phraseological patterns that are directly borrowed (*nice try!* in German), indirectly borrowed (e.g., *der Elephant in Raum* based on *the elephant in the room*) or hybrids characterized by partial substitution of English lexemes by German alternatives (e.g., *den Turnaround schaffen* "to manage the turnaround") (see Andersen, 2020a,b for further cross-linguistic support). This distinction mirrors the opposition made by Matras (2009) between matter and pattern replication. Yet, the construction grammar emphasis on the various degrees of schematicity of constructions becomes crucial when aiming to study not just which phrases are borrowed form a given source language in what form, but also to describe how language users gradually adapt and change these semi-fixed source language chunks in the receptor language. It enables us to integrate, rather than separate, direct, and indirect borrowing, to simultaneously analyze pattern and matter replication, and to arrive at a more profound understanding of the way in which language change takes shape through individual usage occurrences in which language material from source and receptor language are integrated (Boas and Höder, 2018, 2021, and see Traugott and Trousdale, 2013).

Höder (2012) describes just how language users go about such integration (and see Dogruöz and Backus, 2009 for similar ideas earlier). Through interlingual identification, language users conflate similar constructions in two languages as belonging to a common "diaconstruction". When instantiating the diaconstruction in specific usage events, users are left with a choice to lexicalize (parts of) the construction (the pattern) with linguistics elements (the matter) from either language. Whether source or receptor language material is used, may vary or change over time, with language users for instance gradually lexicalizing more slots in the receptor language. This contact-induced change can reflect or support more generally attested types of constructional change (Traugott and Trousdale, 2013; Coussé and Von Mengden, 2014), such as the occurrence of open slots in a previously fully fixed expression, semantic specialization or generalization of the construction as a whole or of its constituent parts.

The benefits of a (Diasystematic) Construction Grammar approach for the description of contact-induced variation and change have been illustrated by the papers in (Boas and Höder, 2018, 2021; and further see for instance Noël and Colleman, 2018). Meanwhile, as concerns change in (semi-)fixed chunks borrowed from English, Zenner et al. (2018) may serve as a first study. In this work, Zenner et al. analyzed semantic and formal similarities between the diaconstructional variants [(DET) ADJ$_{superlative}$ N *ever* (Ptcp)] (e.g., *beste zomer ever* "best summer ever") and [(DET) ADJ$_{superlative}$ N *ooit* (Ptcp)] (e.g., *beste zomer ooit* "best summer ever"). Particularly, Zenner et al. (2018) aimed to link ongoing change in the existing Dutch *ooit*-construction to the incorporation of the English *ever*-counterpart in Dutch. The current paper focuses on a different question. It investigates the construction *Pimp My Ride*, which was borrowed from English into Dutch, and analyzes how its constituent elements have evolved in Dutch both within and outside of the target construction.

Pimp My Ride

In a pilot study, Van de Velde and Zenner (2010) revealed the rapid emergence of the verb *pimpen* "to pimp, fancify" in Dutch following the introduction of *Pimp My Ride* as the name of a popular MTV series in 2005. *Pimp,* part of English vocabulary since the 1600s[3], (1) is a noun referring to a person who controls prostitutes; (2) is a verb derived from the noun in the meanings "to act as a pimp" and "to prostitute someone"; (3) exhibits figurative meanings based on (1) and (2) such as *he pimped himself out to the media*. MTV (Music Television), an American cable channel that spawns numerous affiliated channels across the globe, introduced a then still innovative additional verbalization of the noun when launching the TV show *Pimp My Ride* in 2004[4]. In essence, the show involves the cosmetic makeover of a shabby car provided to the show by a participant. The title of the show refers to the request of the participant (imperative *pimp*) to, simplistically put, fancify their car (*my ride*) to the point where it looks like a car one would stereotypically link to a pimp. Although the show emerged in the socio-cultural context of the US, connecting to cultural ideologies surrounding pimps and ghetto-style, the series, both in the original and in various adapted versions, quickly found an international audience, carrying with it the new meaning of "to pimp".

The choice made by Van de Velde and Zenner (2010) for this particular construction can be justified by pinpointing several reasons why it is actually fairly unexpected that the English phrase *Pimp My Ride* would instigate the introduction of a new verb, *pimpen* "to pimp, fancify" in the Dutch lexicon. First, the verb is introduced in Dutch via an international TV show, and so far the role of media in language change has generally been contested (Tagliamonte, 2014) and, in the context of anglicisms, understudied (Andersen, 2020a, p. 2). Second, its original fairly taboo meaning loaded with social stigma could hinder the spread

---

[3]The *Oxford English Dictionary* includes examples of *pimp,* n. and *pimp,* v. (1) since 1639, and of *pimp,* v. (2) since 1745.
[4]The *Oxford English Dictionary* includes examples of *to pimp* in the meaning "to fancify" from 2000 onwards.

of the more innocent meaning "to fancify". Third, the imperative use of the noun *pimp* concerns a denominal verbalization that is not straightforward in Dutch. Finally, the verb *pimpen* is borrowed as part of a fixed expression functioning as a proper name. Hence, a process of constructional change could be evoked to explain the process of how the verb has become detached from the original proper name.

This latter point is precisely what Van de Velde and Zenner (2010) set out to study. Mining a self-collected fit-for-purpose Dutch corpus of newspaper articles published between 1998 (well before the first episode of *Pimp My Ride* in 2004) and 2009 (the time of analysis) for instances of *pimp*, the authors corroborated the fact that the "fancify" meaning of verbal *pimp* did not occur prior to the introduction of the MTV show *Pimp My Ride* in the Low Countries, in this way pinpointing T0 of the constructional change and sidestepping the actuation problem typical of variational studies (Weinreich et al., 1968). Second, the authors show how, even in the quality newspaper corpus, the new verb *pimpen* with its new "fancify" meaning spread quickly after the introduction of the show *Pimp My Ride*. Making a case for a construction grammatical approach to the change, they show how the verb *pimpen* was distilled from *Pimp My Ride* through a fast process of semantic and formal "deconstructionalization" from the underlying constructional template [*pimp* POSS N]. By "deconstructionalization" we mean the process of a gradual or stepwise blurring of the formal and semantic contours of the constructional template, leading to a less crisply delineated, recognizable chunk. Deconstructionalization is in that sense used here merely as a descriptive label, not as a bid to engage in theoretical and technical terminological debates. We prefer the term "deconstructionalization" over "schematization", as the latter is mainly used to refer to the shift from lexically fixed slots to open slots, whereas we also intend to look at insertion at the syntagmatic level, and expansion of morphological productivity.

Semantically, the verb *pimpen* was increasingly used for other entities than rides and vehicles, even including animates [see (1)][5]. Formally, a transition was noted from the fixed proper name *Pimp My Ride* via the semi-fixed target construction [*pimp* POSS N] [see (2)] to finite uses of the Dutch verb *pimpen* detached from the construction [see (1)].

(1) Vervolgens **pimpten** ze **Frank Deboosere** (*De Morgen,* 12 July 2008).
    "Then they pimped Frank Deboosere [the national broadcaster's weather man]"

(2) onder de slogan **"Pimp uw pots**!" konden vrijwilligers zich een gepimpte muts aanmeten (*HLN*, 14 February 2009).
    "under the slogan "Pimp your hat!" volunteers were able to acquire a pimped hat".

In conclusion, by the end point of the measurements of the pilot study, viz. a mere 5 years after the first broadcast of *Pimp My Ride* in the Low Countries, the (in Dutch) new verb *pimpen* seemed to have established itself in Dutch, being used with a range of pimpable entities, both in the constructional template [*pimp* POSS N] derived from the target expression *pimp My ride* and more freely as a finite verb. This finding is further supported by the inclusion of the verb in dictionaries as of 2006 and the lack of objection to its use in normative reference works[6].

This impressive trajectory in the Low Countries from English-origin TV show to canonized dictionary entry in less than 3 years might hold implications for our understanding of the role of globalized media for local language change. However, as the presumed lack of longevity is a traditional argument used against media-induced variation and change (Labov, 2001, p. 228), a follow-up study is required. Particularly, we aim to understand what has happened with *pimp* in Dutch following the 2004 introduction of *Pimp My Ride*, verifying whether the change is long-lived, and to what extent differences can be found in the trajectories of free use of the new verb *pimpen* "to pimp" and the [*pimp* POSS N]-construction that is closer to and hence in part still resonates the verb's globalized media origin.

Research Questions

This study aims to uncover deconstructionalization patterns in the use of the verb *pimp* in Dutch following the first wave of deconstructionalization as described in Van de Velde and Zenner (2010). A contrast is made between the trajectory of *pimp*-cases in the constructional template [*pimp* POSS N] and of occurrences of verbal *pimp* outside of this original template:

> **RQ1**: To what extent do we find signs of further deconstructionalization in the target construction [*pimp* POSS N] after the establishment of the new verb *pimpen* following the introduction of *Pimp My Ride* in 2004, as attested in the choice of possessive (from original 1SG to other possessors), the type of pimped entity (from vehicles to other entities), the language lexicalizing the pimped entity (from English to Dutch) and the amount of lexical intrusion found in the constructional template (from no intrusion to intrusion)?
>
> **RQ2**: To what extent do we find signs of further deconstructionalization in uses of the verb *pimpen* already detached from [*pimp* POSS N], as attested in verbal morphology (from imperative over infinitive and participle use to finite forms), derivational morphology (from less to more productivity) and in the semantics of the pimped entity (from vehicles over inanimate entities to animate entities)?

METHODOLOGY

To answer the research questions formulated above, we analyzed a sample of 4,561 Dutch tweets derived from a dataset of 163,046 tweets posted between January 2007 and April 2020 including a string of *pimp* (Section Data). Careful manual coding of the tweets for a number of formal and semantic properties (Section Coding Procedure) allowed us to calculate an aggregate score of deconstructionalization both within (RQ1) and outside of (RQ2) the target construction [*pimp* POSS N]

---

[5]Examples in this section are drawn from the database collected by Van de Velde and Zenner (2010).

[6]See https://www.vlaanderen.be/taaladvies/pimpen, https://onzetaal.nl/taaladvies/pimpen, both consulted July 6, 2021.

(Section Deconstructionalization Score). In this way, we go beyond absolute and relative frequency measures, which make up the bulk of the quantitative data in grammaticalization, lexicalization, and constructionalization literature, and combine the attention to diagnostics of constructional change (Traugott and Trousdale, 2013) with a quantitative assessment that is amenable to statistical investigation (see also Petré and Van de Velde, 2018; De Troij and Van de Velde, 2020).

Data

For our study, we made use of a Twitter corpus. Following Androutsopoulos (2010, p. 204), we consider computer-mediated communication (CMC) as optimally suitable when aiming to study how "globalization is not a unidirectional process by which linguistic or cultural elements are diffused and uncritically adopted" but rather a process of local integration. Computer-mediated communication will be a primary source for uncovering the detachment of *pimp* from its original template [*pimp* POSS N]. Practical considerations further support the specific choice for Twitter: it allows us to arrive at a diachronically sliceable corpus of sufficient size. Needless to say, caution is needed when interpreting our findings, in the sense that we cannot simply extrapolate the trajectories of use of *pimp* to other genres or media.

We gathered a dataset of *pimp*-tweets automatically identified as written in Dutch for the period from 2007 to 2020, viz. from the launch of the platform in the Low Countries to the time of data collection. This means that we started our retrieval after the initial wave of deconstructionalization of *Pimp My Ride* in 2004 (Van de Velde and Zenner, 2010). Tweets containing conjugated and derivative forms of *pimp* were gathered through Python, also querying spelling deviations expected to occur in the conjugation of borrowed verb forms in Dutch or in the general CMC context (e.g., *gepimpt* vs. *gepimped*)[7]. All 163,046 collected tweets[8] were lemmatized and POS-tagged with Frog (Van den Bosch et al., 2007). As the quality of POS-tagging is hampered by the multilingual contexts in which many of the *pimp*-forms occur and by the graphemic instability typical of the informality of CMC, we proceeded to manual coding of a sample of tweets. Specifically, a random selection of 500 *pimp*-tweets (or less, if no 500 were available) was selected for each of the 13 years under scrutiny. As such, a total of 6,381 *pimp*-examples was manually coded for their semantic and formal properties.

---

[7]The Twitter API does not allow downloading tweets older than 7 days. We hence relied on the Python package "GetOldTweets3". This package is meanwhile deprecated, with "snscrape" as its successor. As barely any metadata for the tweets is retrieved by GetOldTweets3, we complemented the GetOldTweets3 data with information retrieved from the standard Python package "Tweepy" relying on the tweet's ID.

[8]Twitter API does not guarantee exhaustivity: "Standard search API returns a collection of relevant Tweets matching a specified query. Please note that Twitter's search service and, by extension, the Search API is not meant to be an exhaustive source of Tweets. Not all Tweets will be indexed or made available via the search interface." (https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets).

TABLE 1 | Tokens per observation type.

| Year	pimpMyRide		pimpPOSSN		outsideCx		Total
	N	**%**	**N**	**%**	**N**	**%**	**N**
2007	2	2.13	4	4.25	88	93.62	94
2008	5	2.23	27	12.06	192	85.71	224
2009	12	3.17	43	11.38	323	85.45	378
2010	50	14.93	31	9.25	254	75.82	335
2011	29	8.19	29	8.19	296	83.62	354
2012	51	14.78	34	9.86	260	75.36	345
2013	46	14.11	57	17.49	223	68.40	326
2014	9	2.85	84	26.67	222	70.48	315
2015	10	2.54	81	20.56	303	76.90	394
2016	7	1.89	93	25.14	270	72.97	370
2017	28	6.67	71	16.90	321	76.43	420
2018	4	1.05	69	18.11	308	80.84	381
2019	1	0.30	53	15.73	283	83.97	337
2020	0	0.00	35	12.15	253	87.85	288
Total	254	5.57	711	15.59	3,596	78.84	4,561

Coding Procedure

A first step in the coding procedure consisted of excluding noise from the dataset. A total of 1,820 tweets (28.52%) were excluded from further scrutiny for one of the following reasons: (1) the matrix language of the tweet was not Dutch, but rather English, Afrikaans, German, … ($N = 754$ of $N = 6,381$); (2) the target semantics of *pimp* were not "to fancify", but rather pertained to the original prostitute controlling meaning, or the tweet was too short to establish the meaning of *pimp* ($N = 1,031$ of the remaining $N = 5,627$); (4) the verb *pimp* was conjugated following English derivational rules (e.g., past participle *pimped* rather than *gepimpt*) ($N = 35$ from the remaining $N = 4,596$)[9].

For the $N = 4,561$ observations left after noise removal, we first identified the observation type, contrasting instances within the construction, which include both tokens of the original fixed expression *Pimp My Ride* [$N = 254$; 5.57%, see (3)] and tokens of the derived constructional template [*pimp* POSS N] that are not *Pimp My Ride* [$N = 711$; 15.59%, see (4)], with instances outside of the construction, viz. free occurrences of the verb *pimp* [$N = 3,596$; 78.84%, see (5)]. **Table 1** and **Figure 1** summarize the tokens per observation type per year, supporting the finding that the first wave of deconstructionalization happened soon after the first introduction of *Pimp My Ride* in 2004: the "free" use of *pimp* is the most frequent throughout the period under investigation. Additionally, we see a revival of original *Pimp My Ride* cases in 2010–2013. This probably relates to a rerun of the show on TV.

(3) dat haar van die gozer bij **pimp my ride**. met al die egelstekels.
    "the hair of that dude on pimp my ride. with al those hedgehog spines"

---

[9]The coded data can be accessed via https://osf.io/bjevu/?view_only= 9222b6704473464caf1e242e6c391cdc (folder "deconstructionalization").

FIGURE 1 | Observation type per year.
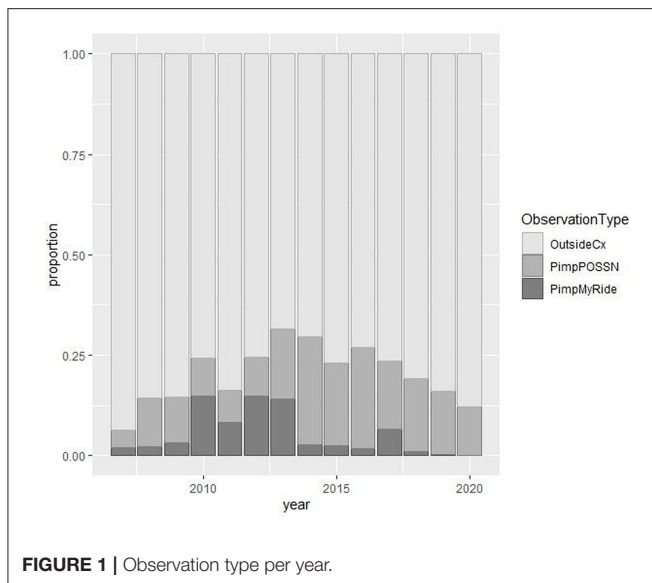
(4) **Pimp my coffee**. Of eigenlijk de koffie van @USERNAME[10].
   "Pimp my coffee. Or actually @USERNAME's coffee."
(5) @USERNAME Hoe zou ge **een chihuahua** dan plastisch willen **pimpen**?
   "@USERNAME How would you want to pimp a chihuahua plastically?"

Next, we proceeded to coding the formal and semantic characteristics for both groups further. **Table 2** summarizes the tokens per year of the construction [*pimp* POSS N], classified according to the parameters of RQ1. This table was composed as follows. Following **Table 1** above, we first split off occurrences of the original fixed proper name *Pimp My Ride*, which always have a 1SG possessive pronoun, *ride* as pimpable entity and English as language for the three slots. For the other [*pimp* POSS N] cases, four parameters are included. First, we indicate whether or not an English first person singular form is used as in the original construction [see *my* in (3) vs. *your* in (6) or *je* "your" in (8)]. For the N-slot, we keep track of the semantics by contrasting pimped vehicles (cars, caravans, motorcycles) with other types of entities [see *car* in (7) vs. *kussensloop* "pillowcase" in (6)]. Further, we keep track of the language used to instantiate N, contrasting English slots with Dutch slots [see *coffee* in (4) vs. *autoband* "car tyre" in (8)][11]. Hybrid forms such as *feestoutfit* "party outfit" are considered English. Unclear cases and proper nouns are awarded NA (*N* = 21, marked red in **Table 2**). Finally, we keep track of lexical elements intruding in the [*pimp* POSS N]-construction [see *nu* "now" in (8)].

---

[10]Personal information of tweets like usernames were replaced by placeholders in the examples.

[11]NAs, marked red in **Table 2**, are awarded in the rare event that the tweet contains an empty placeholder for the N-slot, such as "ZO...... 3E PIMP MY..........! ITEM IS GEMAAKT" *"so... 3rd pimp my ...! item has been created"* (*N* = 4).

(6) Drukke #workshop week 3: Stickeren, Stencil, **Pimp Your Kussensloop**. Guerrilla Gardening, Beatbox, Zang, Theater, Dj en Hiphop op donderdag.
   "Busy #workshop week 3: Stickering, Stencil, Pimp Your Pillowcase. Guerilla Gardening, Beatbox, Singing, Theater, DJ and Hiphop on Thursday."
(7) **Pimp your car** met deze unieke AUTO WIMPERS vandaag bij BRANDNAME.
   "Pimp your car with these uniqe CAR LASHES today at BRANDNAME".
(8) hoe moet ik dat zien? **pimp nu je autoband**?
   "How should I perceive this? Pimp your car tire now?"

The "free" *pimp* tokens do not follow the [*pimp* POSS N]-template, calling for another set of formal parameters that indicate (even) further deconstructionalization. The variables and token counts can be found in **Table 3**. For the semantics, we now resort to a ternary classification, contrasting pimped vehicles, other non-animate pimped entities and animate pimped entities [see (9), (11), and (10)]. NAs are awarded to instances where no pimped entity is specified (*N* = 106).

(9) Zo **de volgbus** van Team Gers! een beetje **op gepimpt** met ballonnen en onze mascotte dog
   "Here the tracking bus of Team Gers! Pimped up a little with balloons and our mascotte dog."
(10) Gistermiddag op de boerderij. Ze **pimpten paardjes** en ook een paar poesjes.
   "Yesterday afternoon on the farm. They pimped horses and a couple of kittens."
(11) #budgettip. **Pimp afdankertjes op**, koop buiten t seizoen en in de #uitverkoop. Maak zo een #cadeaula aan en speel t hele jaar voor Sint.
   "#budgettip. Pimp up discards, buy outside of season and during #sales. Make a #giftdrawer this way and play Santa all year long."

For the formal classification of tokens, we adopt a verbal and a derivational morphological perspective. In terms of verbal morphology, we contrast imperatives [see (11)], infinitives [see (5)], (adjectival use of) participles [see (9)] and finite uses [see (10)]. For derivational morphology, we keep track of *pimp*'s productivity by checking for derivational morphemes or phrasal extension [see *op* "up" in (9)].

Deconstructionalization Score

To arrive at a "holistic" picture of the deconstructionalization process of the lexical expression *pimp my ride* both within and outside the template [*pimp* POSS N], we follow a quantitative procedure applied earlier in studies by Van de Velde (2009, p. 334–339), De Smet and Van de Velde (2013), and Petré and Van de Velde (2018). The idea is that an observation collects "points" for each dimension of constructional change that plays role in the departure of the original construction. To give an example: a point is awarded if we observe *pimp your ride* instead of *pimp my ride*, and yet another if the pimped entity is not *ride* but e.g.,

TABLE 2 | Tokens per category for observations of "pimp my ride" and of the target construction [*pimp* POSS N] (NA's marked in red).

	Pimp My Ride		*pimp* POSS N														N
			POSS				N entity				N lang				Intrusion			
			score = 0 my|ma		score = 1 other		score = 0 vehicle(+NA)		score = 1 other		score = 0 ENG(+NA)		score = 1 NL		score = 0 no		score = 1 yes	
Year	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
2007	2	33.33	2	33.33	2	33.33	1	16.67	3	50.00	4	66.67	0	0.00	4	66.67	0	0.00	6
2008	5	15.63	9	28.13	18	56.25	2	6.25	25	78.13	21	65.63	6	18.75	27	84.38	0	0.00	32
2009	12	21.82	18	32.73	25	45.45	2	3.64	41	74.55	19	34.55	24	43.64	43	78.18	0	0.00	55
2010	50	61.73	8	9.88	23	28.40	3	3.70	28	34.57	20(+1)	25.93	10	12.35	31	38.27	0	0.00	81
2011	29	50.00	10	17.24	19	32.76	4	6.90	25	43.10	14	24.14	15	25.86	28	48.28	1	1.72	58
2012	51	60.00	11	12.94	23	27.06	2	2.35	32	37.65	21	24.71	13	15.29	34	40.00	0	0.00	85
2013	46	44.66	10	9.71	47	45.63	5	4.85	52	50.49	24(+1)	24.27	32	31.07	55	53.40	2	1.94	103
2014	9	9.68	12	12.90	72	77.42	7	7.53	77	82.80	26(+4)	32.26	54	58.06	79	84.95	5	5.38	93
2015	10	10.99	13	14.29	68	74.73	11	12.09	70	76.92	31(+1)	35.16	49	53.85	78	85.71	3	3.30	91
2016	7	7.00	19	19.00	74	74.00	7	7.00	86	86.00	48(+3)	51.00	42	42.00	90	90.00	3	3.00	100
2017	28	28.28	18	18.18	53	53.54	9	9.09	62	62.63	37(+4)	41.41	30	30.30	68	68.69	3	3.03	99
2018	4	5.48	9	12.33	60	82.19	2	2.74	67	91.78	34(+1)	47.95	34	46.58	65	89.04	4	5.48	73
2019	1	1.85	10	18.52	43	79.63	5	9.26	48	88.89	27(+1)	51.85	25	46.30	49	90.74	4	7.41	54
2020	0	0.00	9	25.71	26	74.29	5(+4)	25.71	26	74.29	9(+5)	40.00	21	60.00	35	100.00	0	0.00	35
Total	254		158		553		69		642		356		355		686		25		965

TABLE 3 | Tokens per category for observations outside of the target construction (NA's marked in red).

	Outside Cx																	
	Verbal morphology								Derivational morphology				N entity					
	score = 0 IMP		score = 1 INF		score = 2 (A)PART		score = 3 OTHER		score = 0 NO		score = 1 YES		score = 0 VEH(+NA)		score = 1 OTHER		score = 2 ANIM	
Year	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	Total
2007	0	0.00	30	34.09	51	57.95	7	7.95	88	100.00	0	0.00	8(+3)	12.50	75	85.23	2	2.27	88
2008	2	1.04	97	50.52	85	44.27	8	4.17	189	98.44	3	1.56	16(+5)	10.94	167	86.98	4	2.08	192
2009	6	1.86	131	40.56	158	48.92	28	8.67	313	96.90	10	3.10	48(+12)	18.58	258	79.88	5	1.55	323
2010	2	0.79	122	48.03	101	39.76	29	11.42	252	99.21	2	0.79	34(+11)	17.72	199	78.35	10	3.94	254
2011	2	0.68	144	48.65	129	43.58	21	7.09	287	96.96	9	3.04	32(+9)	13.85	248	83.78	7	2.36	296
2012	7	2.69	125	48.08	113	43.46	15	5.77	242	93.08	18	6.92	31(+7)	14.62	207	79.62	15	5.77	260
2013	5	2.24	95	42.60	103	46.19	20	8.97	208	93.27	15	6.73	29(+6)	15.70	181	81.17	7	3.14	223
2014	16	7.21	95	42.79	90	40.54	21	9.46	212	95.50	10	4.50	31(+6)	16.67	176	79.28	9	4.05	222
2015	14	4.62	117	38.61	117	38.61	55	18.15	287	94.72	16	5.28	62(+8)	23.10	223	73.60	10	3.30	303
2016	16	5.93	140	51.85	91	33.70	23	8.52	253	93.70	17	6.30	53(+7)	22.22	203	75.19	7	2.59	270
2017	20	6.23	169	52.65	84	26.17	48	14.95	315	98.13	6	1.87	100(+5)	32.71	209	65.11	7	2.18	321
2018	14	4.55	148	48.05	114	37.01	32	10.39	296	96.10	12	3.90	79(+9)	28.57	197	63.96	23	7.47	308
2019	12	4.24	122	43.11	121	42.76	28	9.89	260	91.87	23	8.13	47(+11)	20.49	211	74.56	14	4.95	283
2020	6	2.37	118	46.64	109	43.08	20	7.91	231	91.30	22	8.70	57(+7)	25.30	169	66.80	20	7.91	253
Total	122		1,653		1,466		355		3,433		163		733		2,723		140		3,596

pimp your laptop. In a sense, the number of points collected can be seen as a distance value from the original construction[12].

---

[12]NAs systematically receive score 0. Alternative analyses that instead exclude NAs reveal the same results.

The total score is then used as the response variable in a negative binomial regression, with the year of attestation as the predictor. If the predictor is significant, we can assume there to be a robust trend over time. Before we have a look at the results, we will first detail the scoring procedure, which takes
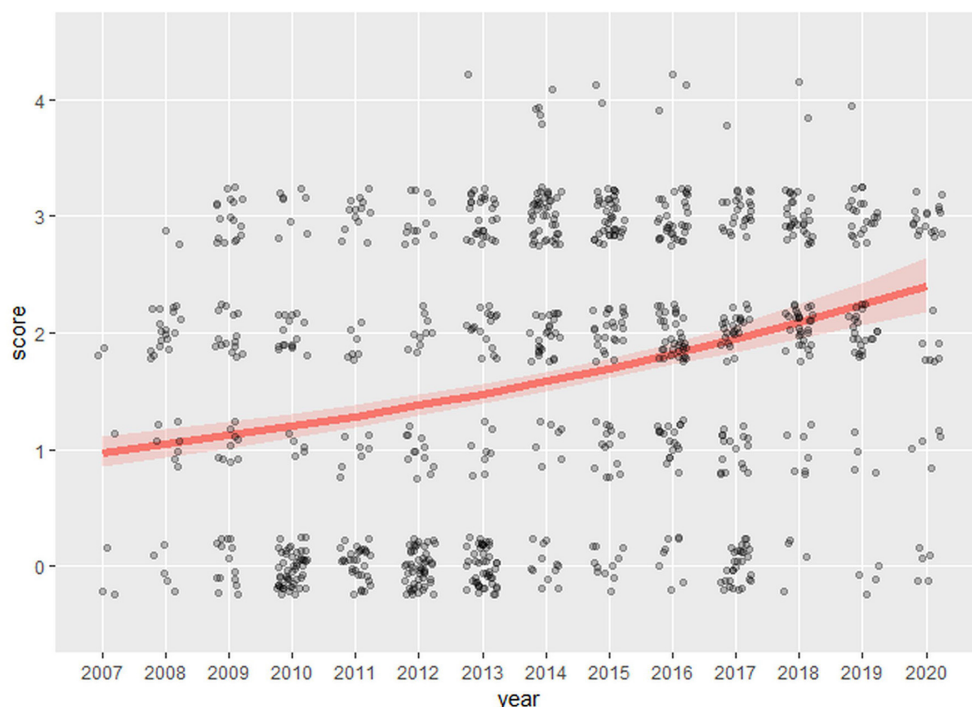
FIGURE 2 | Effect plot negative binomial regression for [pimp POSS N].

the information from the manual coding as input. We made a distinction between the original construction [*pimp* POSS N] and the subsequent "free" use of the verb *pimp*.

For the original construction (RQ1), the following system has been applied:

- Form in the POSS-slot: 0 points if the possessive was *my* or *ma*; +1 point otherwise (*mijn, jouw, onze* ...).
- Semantics of N-slot: 0 points if the entity is a vehicle; +1 point otherwise (food, animate entities etc.).
- Language of the N-slot: 0 if the pimped entity is English, an English loan or a hybrid form; +1 point otherwise (Dutch noun, French loan ...).
- "Intrusion" in the template: 0 if the imperative *pimp*, the possessive, and the pimped entity are contiguously expressed; +1 point if there is intervening material [see *nu* "now" in (8)].

As such, the maximal number of points an observation of the form [*pimp* POSS N] can gather is 4.

For the verb *pimp* in its "free" use, outside of the constructional template (RQ2), the following scoring system has been applied:

- Verbal morphology: 0 if the verb is an imperative; +1 point if the verb is an infinitive; +2 points if the verb is a past participle; +3 points if it is a finite, non-imperative form. This scoring reflects a continuum "infinitive > participle > finite verb": we know from earlier studies that loan verbs enter the (Dutch) language preferably through their non-inflected forms, with infinitives being easier to accommodate

than participles (Wichmann and Wohlgemuth, 2008, see also De Smet, 2014).
- Derivational morphology: 0 if the verb is used as a simple stem; +1 point when signs of productivity are attested, viz. if *pimp* is combined with a derivational morpheme or a phrasal extension (*pimp up, oppimpen, ontpimpen* etc.).
- Entity semantics: 0 if the pimped entity is a vehicle; +1 point if the pimped entity is something else; +2 points if the pimped entity is an animate entity (human, animal or body part).

The maximal number of points for the "free" construction is 6 points.

RESULTS

Patterns of Change Within the Target Construction [*pimp* POSS N]

For the use of the [*pimp* POSS N] construction, the negative binomial model (with the natural logarithm as the link function) indicated a trend over time. Diachronically, the distance from the original construction as measured by the four-point scoring system increases significantly [β = 0.07 (on the log scale), $p <$ 0.001]. This is visually represented in the effect plot in **Figure 2**. Next, **Figure 3** verifies to what extent we still find a significant deconstructionalization trend when taking the original lexical construction, viz. all occurrences of the lexically fixed *Pimp My Ride*, out of the equation. As can be seen in **Figure 3**, the upward trend remains, but loses some of its oomph and its significance [β = 0.01 (log scale), $p = 0.156$].
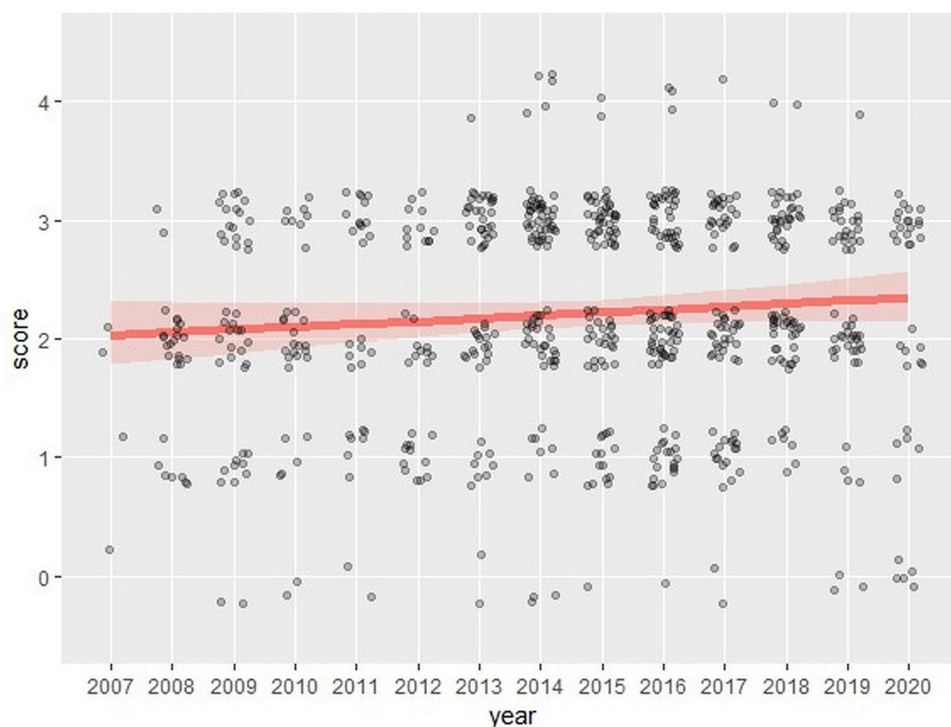
FIGURE 3 | Effect plot negative binomial regression for [pimp POSS N] excluding Pimp My Ride.

Patterns of Change Outside of the Target Construction [*Pimp* POSS N]

"Free-roaming" *pimp*, i.e., the occurrence of *pimp* outside of the [*pimp* POSS N] template, does not show a trend over time. The effect of the year of attestation is not significant [β = −0.004 (log scale), $p = 0.111$]. Indeed, the effect plots in **Figure 4** shows that the line is flat. Upon closer inspection, however, it seems that we would be remiss to assume that everything remains the same. Over time, the average score does remain more or less stable, but the range widens. Indeed, there is an increase over time in the standard deviation (Pearson's correlation 0.83, $p < 0.001$), see **Figure 5**.

What we have, then, is an increase in the higher regions of the scores, but this increase is counterbalanced by a simultaneous upsurge in the more basic use of the construction. This echoes a finding from Zenner et al. (2018), who noted that the use of a new construction can boost the frequency of an older cognate construction, which they call the "a rising tide lifts all the boats" phenomenon.

DISCUSSION AND CONCLUSION

This study aimed to uncover deconstructionalization patterns in the use of the verb *pimp* in Dutch following the first wave of deconstructionalization as described in the Van de Velde and Zenner (2010) pilot study. The analysis distinguished between *pimp*-occurrences found within the

target template [*pimp* POSS N] and of *pimp*-cases that are to be located outside of the original target template. We relied on a holistic score aggregating over formal and semantic diagnostics of the *pimp*-occurrences attested through manual coding.

Within the original constructional template, our aggregate score revealed a significant pattern of deconstructionalization over time. However, the significance attested likely concerns an artifact of a rerun of the show in 2010–2013. This rerun spiked the use of *pimp* in the original fixed phrase *Pimp My Ride*. When this original use started decreasing after 2013, this naturally caused the aggregate deconstructionalization score to rise.

Outside of the construction, not much seems to be going on at first glance. From 2007, our first point of measurement, to 2019, the curve for our deconstructionalization score is flat, indicative of stability in the degree of digression from the original *Pimp My Ride* form and meaning across time. Closer scrutiny revealed a more complicated story. A significant increase is attested in the standard deviations of the aggregate score over time, indicative of an increasingly broad use of *pimp*. This can be understood as the combined effect of two phenomena known from the grammaticalization and constructionalization literature. The first phenomenon is "layering" (see Hopper, 1991): a new use does not overthrow the old use, rather the two may happily coexist. Take for instance the grammaticalization of the indefinite article *a(n)*
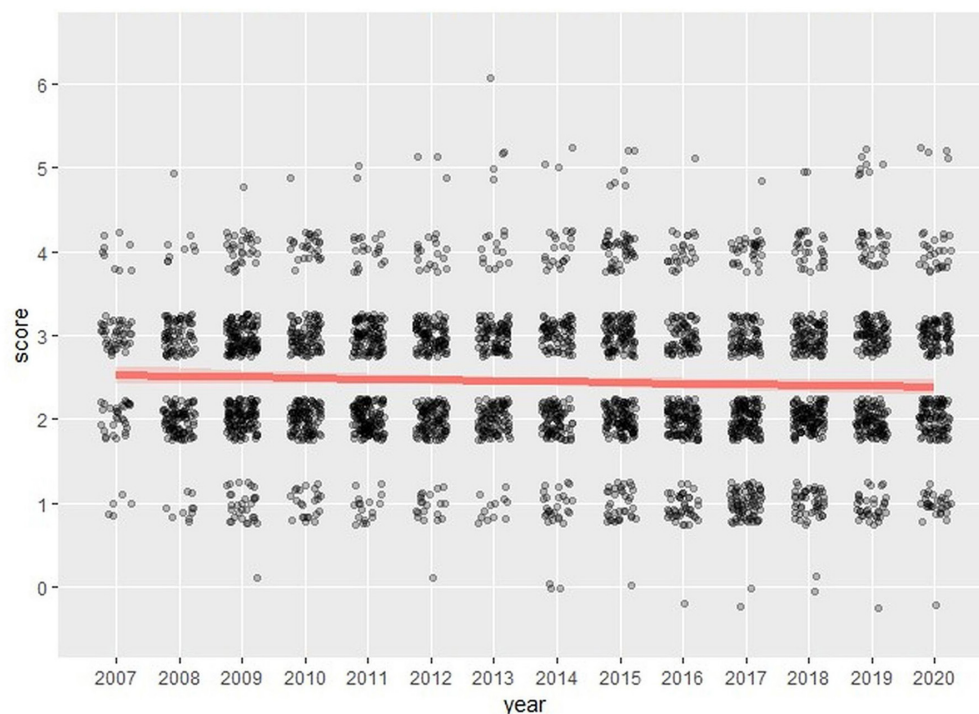
FIGURE 4 | Effect plot negative binomial regression for deconstructionalized "pimp".

from numeral *one*. The emergence of the article does not obliterate the numeral, obviously. The same may be true in our case as well: the emergence of a new use of *pimp* does not necessarily come at the expense of the old fixed expression. The second phenomenon is "a rising tide lifts all the boats" (see Zenner et al., 2018): increased use of the new construction may even lead to a concomitant increase in the use of the old construction.

It would be interesting to complement our corpus study with perception research in at least three ways. First, studies could aim to uncover whether the media origin of this verb is still perceived by language users and to what extent this might promote the selection of the verb over Dutch near-synonyms such as *opleuken*. Second, we could verify to what extent the socio-cultural stereotypes surrounding the pimp persona and ghetto-fabulous style that likely underlie the US original version are perceived by Dutch speakers, or what is "lost in translation". Third, research could verify which of the two *pimp*-meanings ("prostitute controller" or "to fancify") is prompted first in language users" perception and to what extent the negative connotations of the original noun *pimp* might restrict the use of the verb *pimp* or whether instead any trail of negative semantic and social connotation has been bleached from the new verbal usage (see Bucholtz, 2016 on indexical bleaching).

Additionally, to fully grasp the interaction between the global and the local at play in the *pimp* lifecycle, future research can aim to uncover to what extent similar patterns

of deconstructionalization have occurred in other countries where the TV show was broadcast. A quick scan of online dictionaries reveals the occurrence of a *pimp*-verb in our target meaning in German *Duden* (*pimpen*), in Swedish *Akademiens OrdBöcker* (*pimpa*), in Norwegian *NAOB* (*pimpe*) and in English *Cambridge Dictionary* itself (*to pimp*), though not in French *Robert*, in Italian *Treccani* or in Spanish *Diccionario de la lengua española*[13]. It is of course tempting to consider this support for a diasystematic approach to multilingual constructions, as the languages typologically closer to English and hence sharing templates for the noun phrase seem to be the ones who have taken over the construction. Caution is needed, of course, as differences in dubbing or subtitling practices might also be at play, and more or less normative traditions in lexicography might promote or disfavor inclusion of the verb in the dictionary (consider the strong monitoring role of the *Académie française* for French, though see Estival and Pennycook, 2011). Further, the mere occurrence of a *pimp*-based verb does not indicate the extent to which the usage is comparable with the original uses in (American) English. As Andersen (2020a,b) points out, more cross-linguistic research on the way English phrases are included in local languages is needed. As a reviewer of this manuscript rightly points out, such cross-linguistic research

---

[13]https://www.duden.de/rechtschreibung/pimpen, https://svenska.se/, https://naob.no/ordbok, https://dictionary.cambridge.org/dictionary, https://dictionnaire.lerobert.com, https://www.treccani.it/, https://dle.rae.es/, consulted July 7, 2021.
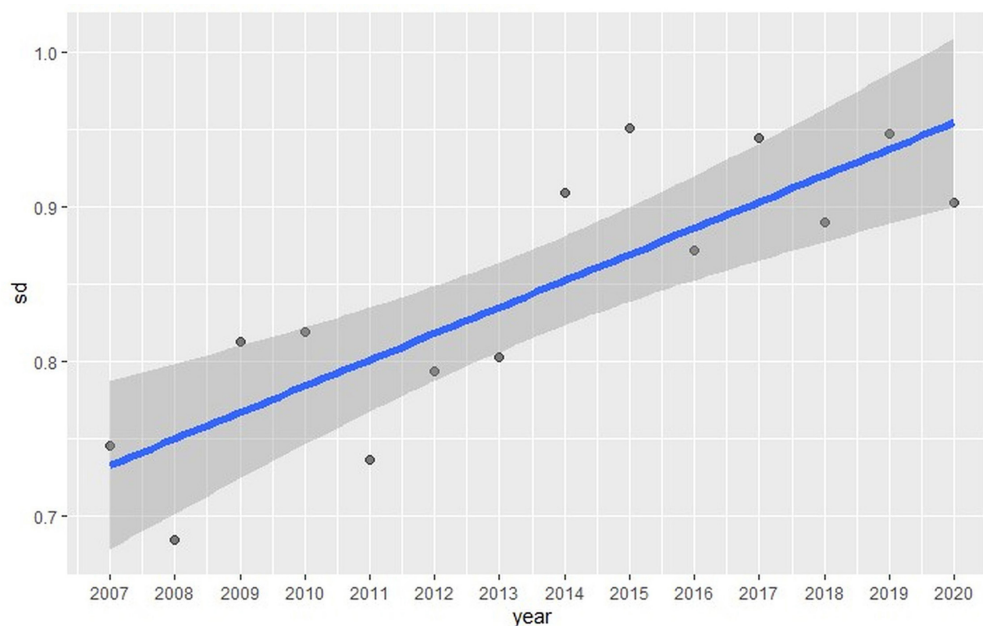
FIGURE 5 | Standard deviations over time.

should also encompass a close comparison with the original usage patterns in the source language. This will allow us to better understand how the global and the local interact, and whether what appears to be local might in itself be more global than anticipated.

Either way, our exploration of the way in which the Dutch-speaking Twitter population has incorporated the new verb *pimpen* in their lexicon points to language users' high flexibility in adopting words from borrowed phrases. On a methodological note, we hope to inspire future work in two ways. First, we hope to reveal the benefits of disentangling free occurrences of *pimp* from occurrences of *pimp* that are part of the constructional imperative template [*pimp* POSS N], and more broadly of identifying points on a spectrum of linguistic innovation from lexically fixed to fully productive patterns. Second, other research might benefit from our scoring system, that allows for a quantified bird's eye perspective derived from manual coding of the formal, morphological and semantic characteristics of the construction at hand.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

SD and DP: conception study, data collection, interpretation results, and revision manuscript. FV and EZ: conception study, data annotation, data analysis, interpretation results, draft and revision manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

Ammon, U. (2001). *The Dominance of English as a Language of Science*. Berlin: Mouton de Gruyter. doi: 10.1515/9783110869484

Andersen, G. (2014). Pragmatic borrowing. *J. Pragmat.* 67, 17–33. doi: 10.1016/j.pragma.2014.03.005

Andersen, G. (2020a). Three cases of phraseological borrowing: a comparative study of as if, Oh wait and the ever construction in the Scandinavian languages. *Ampersand* 7, 100062. doi: 10.1016/j.amper.2020.100062

Andersen, G. (2020b). "Phraseology in a cross-linguistic perspective: a diachronic and corpus-based account," in *Corpus Linguistics and Linguistic Theory*. doi: 10.1515/cllt-2019-0057

Androutsopoulos, J. (2010). "Localizing the global on the participatory web," in *The Handbook of Language and Globalization*, ed N. Coupland (London: Blackwell), 201–231. doi: 10.1002/9781444324068.ch9

Blommaert, J. (2010). *The Sociolinguistics of Globalization*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511845307

Boas, H. (2013). "Cognitive construction grammar," in *The Oxford Handbook of Construction Grammar,* eds T. Hoffmann and G. Trousdale (Oxford: Oxford University Press), 233–154. doi: 10.1093/oxfordhb/9780195396683.013.0013

Boas, H.C., and Höder, S. (eds.). (2018). *Constructions in Contact*. Amsterdam: Benjamins. doi: 10.1075/cal.24

Boas, H.C., and Höder, S. (eds.). (2021). *Constructions in Contact 2. Language Change, Multilingual Practices, and Additional Language Acquisition.* Amsterdam: Benjamins. doi: 10.1075/cal.30

Bucholtz, M. (2016). "On being called out of one's name: indexical bleaching as a technique of deracialization," in *Raciolinguistics: How Language Shapes Our Ideas About Race,* eds H. Samy Alim, John R. Rickford, and Arnetha F. Ball (Oxford: OUP), 273–289. doi: 10.1093/acprof:oso/9780190625696.003.0016

Coussé, E., and Von Mengden, F. (eds). (2014). *Usage-Based Approaches to Language Change.* Amsterdam: Benjamins. doi: 10.1075/sfsl.69

Cowie, A. (2001). *Phraseology: Theory, Analysis, and Applications.* Oxford: OUP.

Croft, W., and Cruse, A. D. (2004). *Cognitive Linguistics.* Cambridge: CUP. doi: 10.1017/CBO9780511803864

De Smet, H. (2014). "De integratie van Engelse leenwerkwoorden in het Nederlands," in *Patroon en argument: Een dubbelfeestbundel bij het emeritaat van William Van Belle en Joop van der Horst,* eds F. Van de Velde, H. Smessaert, F. Van Eynde, and S. Verbrugge (Leuven: Leuven University Press), 75–87. doi: 10.2307/j.ctt14jxsr0.8

De Smet, H., and Van de Velde, F. (2013). Serving two masters: form-function friction in language change. *Stud. Lang.* 37, 534–565. doi: 10.1075/sl.37.3.04des

De Troij, R., and Van de Velde, F. (2020). Beyond mere text frequency: assessing subtle grammaticalization by different quantitative measures. A case study on the Dutch soort construction. *Languages* 55, 1–17. doi: 10.3390/languages5040055

Dogruöz, S., and Backus, A. (2009). Innovative constructions in Dutch Turkish: an assessment of ongoing contact-induced change. *Bilingual. Lang. Cogn.* 12, 41–63. doi: 10.1017/S1366728908003441

Estival, D., and Pennycook, A. (2011). L'Académie française and Anglophone language ideologies. *Lang. Policy* 10, 325–341. doi: 10.1007/s10993-011-9215-6

Feyaerts, K. (2006). "Phraseology of Dutch," in *Phraseologie/Phraseology,* eds H. Burger, D. Dobrovol'skij, P. Kühn and N. R. Norrick (Berlin: Mouton De Gruyter), 644–654.

Fiedler, S. (2017). Phraseological borrowing from English into German: cultural and pragmatic implications. *J. Pragmat.* 113, 89.e102. doi: 10.1016/j.pragma.2017.03.002

Fillmore, C. J. (1988). "The mechanisms of "construction grammar"," in *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society.* UC Berkeley, 35–55. doi: 10.3765/bls.v14i0.1794

Garley, M. (2018). "Peaze up! Adaptation, innovation, and variation in german hip hop discourse," in *Multilingual Youth Practices in Computer Mediated Communication,* eds C. Cutler and U. Royneland (Cambridge: CUP), 87–108. doi: 10.1017/9781316135570.006

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure.* Chicago, IL: University of Chicago Press.

Goldberg, A. (2006). *Constructions at Work: the Nature of Generalization in Language.* Oxford: Oxford University Press.

Goldberg, A. E. (1999). "The emergence of the semantics of argument structure constructions," in *Emergence of Language,* ed B. Macwhinney (Hillsdale, NJ: Lawrence Earlbaum Associates), 197–212.

Görlach, M. (2003). *English Words Abroad.* Amsterdam; Philadelphia, PA: John Benjamins. doi: 10.1075/tlrp.7

Granger, S. A. (2009). "Window onto the L2 phrasicon," in *Collocating in another language: Multiple interpretations,* eds A. Barfield and H. Gyllstad (Palgrave Macmillan), 60–65.

Gries, S. (2008). "Phraseology and linguistic theory: a brief survey," in *Phraseology: An Interdisciplinary Perspective,* eds S. Granger and F. Meunier (Amsterdam: John Benjamins), 3–26. doi: 10.1075/z.139.06gri

Höder, S. (2012). "Multilingual constructions: a diasystematic approach to common structures," in *Multilingual Individuals and Multilingual Societies,* eds K.Braunmüller and C. Gabriel, (Amsterdam: Benjamins), 241–257. doi: 10.1075/hsm.13.17hod

Hopper, P. J. (1991). "On some principles of grammaticalization," in *Approaches to Grammaticalization. Volume I: Focus on Theoretical and Methodological Issues,* eds E. C. Traugott and B. Heine (Amsterdam: John Benjamins),17–35. doi: 10.1075/tsl.19.1.04hop

Kay, P., and Michaelis, L. (2011). "Constructional meaning and compositionality," in *Semantics: An International Handbook of Natural Language Meaning, Vol.3, HSK Handbooks of Linguistics and Communication Science,* eds C. Maienborn, K. von Heusinger, and P. Portner (Berlin: Mouton de Gruyter), 2271–2296.

Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors.* Oxford: Blackwell.

Leppänen, S. (2007). Youth language in media contexts: insights into the functions of English in Finland. *World Englishes* 26, 149–169. doi: 10.1111/j.1467-971X.2007.00499.x

Matras, Y. (2009). *Language Contact.* Cambridge: CUP. doi: 10.1017/CBO9780511809873

Mauranen, A., and Ranta, E. (2009). *English as a Lingua Franca: Studies and Findings.* Newcastle upon Tyne: Cambridge Scholars Press.

Michaelis, L. (2019). "Constructions are patterns and so are fixed expressions," in *Patterns in Language and Linguistics,* eds B. Busse and R. Moehlig-Falke (Berlin: Mouton), 193–220. doi: 10.1515/9783110596656-008

Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes.* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198299530.001.0001

Noël, D., and Colleman, T. (2018). "Constructional renovation : the role of French legal language in the survival of the nominative-and-infinitive in Dutch," in *Cognitive Contact Linguistics : Placing Usage, Meaning and Mind at the Core of Contact-Induced Variation And Change,* eds E. Zenner, A. Backus and E. Winter-Froemel (Berlin: De Gruyter), 305–338. doi: 10.1515/9783110619430-011

Onysko, A. (2007). *Anglicisms in German. Borrowing, Lexical Productivity and Written Codeswitching.* Berlin; New York, NY: Walter de Gruyter. doi: 10.1515/9783110912173

Onysko, A. (2021). "An integrated perspective of diasystematic construction grammar and cognitive language contact phenotypes," in *Constructions in Contact 2. Language Change, Multilingual Practices, and Additional Language Acquisition,* eds H. C. Boas and S. Höder (Amsterdam: John Benjamins), 82–105. doi: 10.1075/cal.30.04ony

Onysko, A., and Winter-Froemel, E. (2011). Necessary loans - luxury loans? Exploring the pragmatic dimension of borrowing. *J. Pragmat.* 43, 1550–1567. doi: 10.1016/j.pragma.2010.12.004

Peterson, E., and Beers Fägersten, K. (2018). Introduction to the special issue: linguistic and pragmatic outcomes of contact with English. *J. Pragmat.* 133, 105–108. doi: 10.1016/j.pragma.2018.06.005

Petré, P., and Van de Velde, F. (2018). The real-time dynamics of the individual and the community in grammaticalization. *Language* 94, 867–901. doi: 10.1353/lan.2018.0056

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. *Linguistics* 18, 581–618. doi: 10.1515/ling.1980.18.7-8.581

Posthumus, J. (1986). *A Description of a Corpus of Anglicisms.* Groningen: Anglicistisch Instituut.

Sawada, S. (2000). The semantics of the "Body Part Off" Construction. *English Linguist.* 17, 361–385. doi: 10.9793/elsj1984.17.361

Steels, L. (ed.). (2011). *Introducing Fluid Construction Grammar. Design Patterns in Fluid Construction Grammar.* Amsterdam: John Benjamins. doi: 10.1075/cal.11.03ste

Tagliamonte, S. (2014). Situating media influence in sociolinguistic context. *J. Sociolinguist.* 18, 223–232. doi: 10.1111/josl.12074

Traugott, E. C., and Trousdale, G. (2013). *Constructionalization and Constructional Change.* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199679898.001.0001

Vaattovaara, J., and Peterson, E. (2019). Same old paska or new shit? On the stylistic boundaries and social meaning potentials of swearing loanwords in Finnish. *Ampersand* 6. doi: 10.1016/j.amper.2019.100057

Van de Velde, F. (2009). *De nominale constituent. Structuur en geschiedenis.* Leuven: Leuven University Press.

Van de Velde, F., and Zenner, E. (2010). "Pimp my Lexis: het nut van corpusonderzoek in normatief taaladvies," in *Liever meer of juist minder? Over normen en variatie in taal,* eds E. Hendrickx, K. Hendrickx, W. Martin, H. Smessaert, W. Van Belle, and J. Van der Horst (Gent: Academia Press), 51–68.

Van den Bosch, A., Busser, B. J., Daelemans, W., and Canisius, S. (2007). "An efficient memory-based morphosyntactic tagger and parser for Dutch." in *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting,* eds F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Leuven, Belgium), 99–114.

Van Meurs, F. (2010). *English in Job Advertisements in the Netherlands - Reasons, Use and Effects*. Utrecht: LOT.

Weinreich, U., Labov, W., and Herzog, M. (1968). "Empirical foundations for a theory of language change," in *Directions for Historical Linguistics: A Symposium*, eds W. P. Lehmann and Y. Malkiel (Austin: University of Texas Press), 95–195.

Wichmann, S., and Wohlgemuth, J. (2008). "Loan verbs in a typological perspective," in *Aspects of Language Contact: New Theoretical, Methodological and Empirical Findings With Special Focus on Romancisation Processes*, eds T. Stolz, D. Bakker and R. Salas Palomo (Berlin: Mouton de Gruyter), 89–121.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: CUP. doi: 10.1017/CBO9780511519772

Yang, W. (1990). *Anglizismen im Deutschen*. Tübingen: Max Niemeyer Verlag. doi: 10.1515/9783111676159

Zenner, E., Backus, A., and Winter-Froemel, E. (eds.). (2019). *Cognitive Contact Linguistics*. Berlin : Mouton de Gruyter. doi: 10.1515/9783110619430

Zenner, E., and Geeraerts, D. (2015). I'm queen of the world! (Semi-)fixed English expressions and constructions in Dutch. *Taal Tongval* 67, 247–274. doi: 10.5117/TET2015.2.ZENN

Zenner, E., Heylen, H., and Van de Velde, F. (2018). Most borrowable construction ever! A large-scale approach to contact-induced pragmatic change. *J. Pragmat.* 133, 134–149. doi: 10.1016/j.pragma.2018.05.004

Zenner, E., Speelman, D., and Geeraerts, D. (2014). A sociolinguistic analysis of borrowing in weak contact situations: english loanwords and phrases in expressive utterances in a Dutch reality TV show. *Int. J. Bilingual.* 19, 333–346. doi: 10.1177/1367006914521699

Ziem, A. (2018). Construction grammar meets phraseology: eine standortbestimmung. *Linguistik* 90. doi: 10.13092/lo.90.4316

frontiers | Frontiers in Communication

Global Englishes and the Semiotics of German Radio—Encouraging the Listener's Visual Imagination Through Translingual and Transmodal Practices

Sarah Josefine Schaefer*

School of Cultures, Languages and Area Studies, University of Nottingham, Nottingham, United Kingdom

Due to increasing mediatisation of societies and the global diffusion of English, previous research has paid a great deal of attention to the use of English in contact with other linguistic resources. Traditionally, the focus in studies on global Englishes is predominantly on verbal resources, which are often examined in media corpora. The recent paradigmatic shift towards a conceptualisation of language as a social practice, however, also acknowledges other resources such as music, images and sounds as part of semiotic assemblages in the process of meaning-making. This paper contributes to this current debate and argues for a more holistic view on language illustrated by examples of the use of global Englishes in German radio media texts. The examples show the complexities of translingual and transmodal practices in mass media communication and how English linguistic resources are deeply entangled with other semiotic resources and thereby locally appropriated in semiotic assemblages to stimulate the listener's visual imagination and achieve communicative success in a non-visual medium.

Keywords: global Englishes, German, transmodality, translanguaging, radio, media, semiotics, mobility

INTRODUCTION

When people think about language, they usually think of words and sentences in the form of oral conversations or written texts. This prioritisation of the verbal aspects of communication is also present in the study of English worldwide. The English language has been the centre of attention in the field of World Englishes, which is mainly interested in the spread and development of varieties of English across the globe (amongst others Kachru, 1985; Schneider, 2007; Mair, 2013; Onysko, 2016). In relation to the global dissemination of English linguistic resources, there is also research on what is frequently called anglicisms, namely English words and phrases that are used in many domains in localities of the expanding circle (Kachru, 1985) where English does not hold an official status, such as in France, Italy, Spain and Germany (Picone, 1996; Onysko, 2007; Furiassi, 2010; Pulcini et al., 2012; Andersen, 2015).

Recently, however, the predominant focus on verbal resources in the study of English worldwide has come under criticism since, when we take a closer look at communication, there is a lot more to say than what the focus on verbal language is able to reveal (see Pennycook, 2017, 2020; Canagarajah, 2018a; Li, 2018). There are, for example, sounds, silence, music and gestures that are language too since they are part of the semiotic practices we engage in. These modes, as they

are called, are "socially shaped and culturally given semiotic resource[s]" (Kress, 2010, p. 79) and contribute a great deal to how we make meaning of the world and how we communicate it. According to Canagarajah, "these broader semiotic resources not only account for the meaning of words in an interaction, they themselves contribute meanings that need to be taken into account in order to explain intelligibility and communicative success in specific interactions" (2018a, p. 806).

Researchers in critical sociolinguistics have therefore drawn on the notion of multimodality as developed in social semiotics and related fields (amongst others Van Leeuwen, 1999; Bateman, 2008; Kress, 2010; Burn, 2014; Stöckl, 2016) to facilitate an analysis of meaning-making that goes beyond verbal resources in linguistic analyses and have at the same time developed this concept further. Rather than viewing meaning-making as a compilation of several, separate modes or independent channels of meaning, as implied by the term "multimodal" (Van Leeuwen, 1999; Kress, 2010), Pennycook (2007), for example, has proposed the term *transmodality* to describe the fact that verbal language does never occur in isolation but is always part of an assemblage of several semiotic modes. In addition, unlike the concept of multimodality, the notion of transmodality is rooted in the perspective that Blommaert (2010) has referred to as the sociolinguistics of mobility. In this paradigm, language is viewed as consisting of mobile linguistic resources, and the boundaries between individual "languages" are considered as socially and politically constructed. A key argument within this perspective is that we have to overcome the structuralist legacy that is inherently part of concepts such as "multilingualism" or "varieties of English," which merely multiply the number of bounded "languages" and thereby reify the monolithic concept of language they originally have set out to supersede (cf. Kachru, 1985). The concept of transmodality therefore complements the critical sociolinguistic concepts of translingualism (Canagarajah, 2013) and translanguaging (Creese and Blackledge, 2011; Otheguy et al., 2015; Li, 2018), which state that people make use of a repertoire of various semiotic resources rather than of separate languages or varieties thereof. Such a holistic perspective on language also requires a re-evaluation of the traditional distinction between text and context since "features we may have treated as part of context may constitute an assemblage that is integral to meanings and communication" (Canagarajah, 2018b, p. 34). Taken together these critical concepts question the separability of different modes of meaning-making and call for a reconceptualisation of language as a trans-lingual, trans-semiotic and trans-modal practice. Against this theoretical background, Canagarajah (2013) and Pennycook (2007) use the term *global Englishes* to highlight the need for focussing on local social and cultural practices associated with the worldwide use of mobile Englishes.

In the light of these recent developments around the conceptualisation of language, this paper follows Canagarajah's (2018a) call to broaden the scope of analysis in research on the worldwide use of English by including other modes besides speech and writing but also by acknowledging the interaction between English and other linguistic resources in analyses of meaning-making. Taking Canagarajah's (2018a) example of the use of English in interpersonal workplace communication between non-native speakers of English as a point of departure, I will show that a translingual and transmodal perspective can also help us to better understand how English linguistic resources are embedded in the semiotic assemblages of representations of events in mass media messages. While recent studies on global Englishes in social media have already adopted translanguaging and transmodality approaches (Sharma, 2012; Sultana and Dovchin, 2017; Baker and Sangiamchit, 2019), the way journalists use English linguistic resources as part of transmodal assemblages to achieve communicative success in traditional mass media messages has been largely overlooked in World Englishes and anglicism research (amongst others Yang, 1990; Glahn, 2002; Mesthrie, 2002; Adler, 2004; Zenner et al., 2012; Makalela, 2013; Lee, 2014; Gottlieb, 2015). I will focus on German radio media in this paper since Germany is one of the localities in the expanding circle where present-day dynamics of mobile Englishes can be observed (Schneider, 2014; Onysko, 2016). The German radio media are not only affected by global cultural flows and the associated diffusion of English but are also part of the dissemination processes of new words, mostly in the form of anglicisms in Germany, as well as of Anglo-American popular culture including music, trends and technological advances (Schaefer, 2019, 2021a,b).

To compensate for radio's lack of visual elements, radio journalists rely on several auditory modes to produce meaning, which are orchestrated in a complex acoustic mix of mainly speech, music and sounds to grab the listeners' attention and to emulate a multisensory effect (Crisell, 1986; Shingler and Wieringa, 1998; Miller, 2018). People listen to the radio in many everyday contexts, such as at home, whilst driving or at the shopping centre. Therefore, radio functions as a background medium, and the components of a radio broadcast and how the medium's message is constructed to achieve communicative success often go unnoticed. I will show that this auditory-only medium reveals complex, transmodally composed meanings to stimulate the audience's imagination and to enable the listener to "visualise" events reported on by journalists. These transmodally composed meanings are worthy of our attention to improve our understanding of the functions of English linguistic resources and their local appropriation through their entanglement with other semiotic resources in mass-mediated communication. For my examples of radio texts, I draw on a large adult contemporary[1] radio corpus containing 60 h of radio morning show content, which I compiled in 2016 as part of a larger, ongoing research project on the use of English linguistic resources by German radio journalists. While I focus on the embeddedness of English linguistic resources in semiotic assemblages in the two examples below, there are also other factors that shape the use of global Englishes by German radio journalists, which I have elaborated on in previous studies. These include broadly speaking, the individual journalist (including his/her language perceptions) (see Schaefer, 2021a), the media organisation (including radio format, genre and domain) (see Schaefer, 2019, 2021a), the

---

[1] Adult contemporary stations target an audience between 25 and 49 years of age and play mainstream pop-music.

Time (sec)	Speech	Music	Sounds
43.3–45.9	**C (studio):** Aber alle haben nur auf ihn hier gewartet. [But everyone has just been waiting for him.]	Hip-hop music, live, French lyrics	Ambient noises of crowd
46.0–50.0	**F:** *David Guetta. Sa voix.* [His voice.] *David.* **Wow!**		
50.1–52.5	**C (concert): I go crazy.** Ich werd' verrückt. [I go crazy.] David Guetta.		
52.6–54.8	**F: David Guetta, the real David Guetta. Uh!**	Concert live music played in background	
54.9–63.0	**C (studio):** David Guettas Heimspiel, der DJ ist ja gebürtiger Pariser, war eine Sensation. Unter dem Eiffelturm auftreten zu dürfen, das machte ihn schwer glücklich. [David Guetta's home match, the DJ is a native Parisian, was a sensation. To be able to perform beneath the Eiffel Tower made him really happy.]		

C correspondent (in studio *or at* concert*); F French concertgoer.*
German: standard font; English: bold; French: italics.
English translations in square brackets.

broadcasting system (public service vs. private radio stations) and the global/cultural background (see Schaefer, 2021b).

SEMIOTIC ASSEMBLAGES ON GERMAN RADIO

The first example from German radio that highlights the complex interplay of modes and the role of English in this context is a foreign correspondent's report about a concert which took place in the context of the 2016 European football championship in France. The excerpt shown in **Table 1** is taken from a report (90 s in length) in which the correspondent shares his experiences of the concert held in the fan area beneath the Eiffel Tower with the radio station's audience back in Germany. The report begins with a description of the security checks necessary for concertgoers to be admitted to the actual event, which contains several French utterances played in the background and a French direct quote of a concertgoer (translated into German by the correspondent). These elements set the scene of the large-scale event at the Eiffel Tower and are used by the journalist to create authenticity. The excerpt from the journalistic piece in **Table 1** shows how the concert audience has experienced the highlight of the event, the performance of the French DJ David Guetta.

As becomes evident from this example, the modes of sound, speech and music interact to create an authentic representation of the concert happenings and to allow the listener to generate a visual image of the event. The report consists of narrative parts recorded by the journalist in studio and of on-scene recordings, which becomes evident to the listener through the change in sound characteristics between the portable recorder and the in-studio recording equipment. Throughout the report, the atmosphere on location is upheld by ambient noises of the large crowd at the concert and at times live music played in the background. The correspondent takes on two distinct personae in the report, which are also mirrored in his linguistic choices. Through the voiceover narration recorded in studio, he acts as an omniscient narrator, who addresses his radio audience and comments on the happenings from an observer's perspective using past tense.

On location, however, the reporter's role is more complex. In contrast to his factual tone in the omniscient narration in studio, the foreign correspondent's word choice and voice at the scene mirror the international setting and the cheering atmosphere at the concert. His excitement for the appearance of the DJ is expressed by the tone of his voice when he shouts "I go crazy. Ich werd' verrückt. David Guetta." On location, the correspondent is part of the crowd, therefore, a participant at the event interacting with other concertgoers. This becomes evident from non-verbal features such as the loudness and clarity of the voices of both the journalist and a French concertgoer, in contrast to the ambient sound of the crowd cheering at the concert, implying that the two are interlocutors standing next to each other, addressing each other. In their short conversation, the German correspondent and the French spectator perform translanguaging by means of the two interlocutors making use of their expanded language repertoire in the form of English resources. In addition, what is interesting to note here is that the utterance "I go crazy. Ich werd' verrückt. David Guetta" by the journalist indicates that his message is not only a response to the French concertgoer's exclamation, but that the translation of *I go crazy* by the correspondent is additionally directed at the radio audience at home in Germany, allowing the listener to become part of the excitement shared by the two concertgoers in conversation. The journalist's interactions at the scene therefore connect several communicative levels. In addition to his role as a participant of an international event on location, the correspondent moves into the role of the reporter who caters for his radio audience's linguistic needs in terms of comprehensibility on a translocal level. In line with the reporter's dual role on location, the English phrase *I go crazy* also has a dual function. On the one hand, it is part of his translingual practice with the French concertgoer, where the journalist makes use of shared English resources to facilitate communication with his immediate surroundings at the concert. On the other hand, it functions to symbolise the international character of the event (in conjunction with the French elements) for the radio station's audience. This dual

TABLE 2 | Promotion piece.

Time (sec)	Speech	Music	Sounds	
0.0–1.8	**Station voice:** Dieses Wochenende ist es... [This weekend it is...]	Rock music (light electric guitar riffs)		
1.9–3.8	... wieder so weit. [... time again.]		Motorcycle start-up; Idling engine	
3.9–7.1	Auf geht's zum großen Motorradwochenende... [Off to the big motorcycle weekend...]			
7.2–13.2	... in X. Der offizielle Start in die Bike-Saison 2016. [... in X. The official start into the biking season 2016.]	(final chord)		
13.3–15.6	Am Samstag Motorradmarkt,... [On Saturday, motorcycle market,...]	Rock music (electric guitar solo, heavy distortion)		
15.7–18.0	... Bike-Shows, großes Familienprogramm... [... bike shows, large family programme...]		Cheering crowd	Revving motorcycle engine
18.1–24.2	... und Live-Musik. Und Sonntag, die traditionelle Motorradsternfahrt mit tausenden Bikes... [...and live music. And on Sunday, the traditional motorcycle rally with thousands of bikes...]			
24.3–25.5	... aus ganz X. [... from all over X.]		Motorcycle passing	

English translations in square brackets. The pseudonym X is used for the radio station's location for anonymisation purposes.

function of the English phrase only becomes evident through the transmodal assemblage of verbal resources (i.e., the translation for the German radio audience) and non-verbal features, as mentioned above, creating an image of spatial proximity between the two concertgoers.

This example reveals a complex orchestration of English linguistic resources and other semiotic resources that acoustically reconstructs the highlight of the international event for the German target audience, stimulating the listener's visual imagination of the event and thereby connecting the local radio audience to a more global context. It additionally demonstrates that taking a closer look at these transmodal representations of events in mass-mediated communication allows us to reveal complex communicative functions of English linguistic resources deeply entangled in semiotic ensembles that contribute to achieving communicative success.

The second example from German radio shown in **Table 2** is a promotion piece about an event sponsored by a radio station in which English loanwords are used. I will focus on the function of the English linguistic resource *Bike*, appearing also in the compound nouns *Bike-Saison* "biking season" and *Bike-Show* in the promotion piece, and how it is embedded in the ensemble of semiotic modes. When we consider the lexical meaning of *Bike*, the loanword is polysemous in German and can refer to different classes of two-wheeled vehicles, namely bicycles and motorcycles[2]. More specifically, it can be used to refer to various types of these vehicles, such as mountain bikes, pedelecs, superbikes and cruiser motorbikes. The actual meaning of the term in this promotion piece and therefore the type of event

advertised by the radio station only becomes evident through the transmodal assemblage it is used in.

Let us first look at some further clues on the verbal level. The use of the German word *Motorrad* "motorcycle" as modifier of the compound noun *Motorradwochenende* "motorcycle weekend" in the promotion piece gives some indications that the event will centre around motorised vehicles, instead of bicycles. However, even in combination the English and German linguistic resources only give a rough idea of what the event is about since, in this sense, they are both hypernyms for a class of vehicles, which still leaves uncertainty regarding the type of motorcycles the event centres around. The modes of sound and music close this gap by creating an acoustic image, a soundscape so to speak, of the event. The engine sounds accompanying the promotion conjure up images of heavy choppers or cruisers which evoke stereotypes associated with the American motorcycle brand Harley Davidson or similar brands. This imagery is supported by the rock music played in the background, which together with the engine sounds rather conjures up German clichés of American motorcycle ideology often related to US-specific cultural concepts such as the Route 66 and the motion picture Easy Rider, rather than to the Moto GP racing series. The English linguistic resource *Bike*, therefore, functions in orchestration with the stereotypic engine sounds and the rock music played in the background to represent American biker ideology and lifestyle. This imagery conveyed at the opening of the promotion piece is additionally strengthened by the term *Bike-Show*, which in this context relates the event to a particular type of motorcycle competition originating in the USA.

What is interesting to note here, however, is that the sounds of heavy chopper or cruiser bike engines and the rock music played together with the information provided verbally to this point may lead to a misinterpretation by the target audience regarding the type of event promoted. The event is seemingly portrayed as a meeting of this particular biker community

---

[2]see 'Bike', in: DWDS – Digitales Wörterbuch der deutschen Sprache, edited by Berlin-Brandenburgische Akademie der Wissenschaften. Available online at: https://www.dwds.de/wb/Bike (accessed June 20, 2021).

only or at best motorcycle enthusiasts who share an interest in American motorcycle culture more generally. To avoid such unintended interpretations on behalf of the listener, the station has included additional verbal clues by using the phrase *großes Familienprogramm* "large family programme," which reframes the meaning of the promotion piece and the meaning of the semiotic resources used to this point by portraying an event that everyone who wants to have a good time can attend.

As this example shows, the actual function of the English word *Bike* in this piece only becomes evident from the interplay of the different modes of speech, music and sounds used to acoustically represent this social event, which together provide a description of the American atmosphere that the station's audience can expect to find when attending this particular event. This semiotic ensemble, however, becomes reframed onto a local scale by making clear on a verbal level through other linguistic resources that this is an event adapted to a wide and at the same time local audience. The meaning of the English loanword *Bike* in this context is therefore co-constructed through its embeddedness in an assemblage of the modes of music, speech and sound, which together kindle a visual imagery of the event with the listener.

THE NEED FOR A CRITICAL PERSPECTIVE ON MASS MEDIA LANGUAGE

As my examples from German radio content have shown, a holistic, translingual and transmodal approach acknowledging semiotic assemblages can widen our understanding of the complex entanglements of English linguistic resources with other semiotic resources and thereby of their local appropriation in mass media content. In line with Canagarajah's (2018a) observations, the examples discussed in this paper have furthermore revealed how these diverse semiotic resources are

used by journalists on radio to achieve communicative success. Sounds, music and speech all contribute meaning in human interaction. Sounds and music can co-construct the meaning of a verbal resource, which shows that other semiotic resources besides speech are not simply part of the context but an intrinsic part of translingual practices. Furthermore, we have seen that various semiotic resources are used on radio to substitute the missing visual elements by creating acoustic representations of events stimulating the listener's imagination. As stated by Canagarajah (2018a), Li (2018), and Pennycook (2017, 2020), in times of increased mobility of cultural and linguistic resources across space and time, it is necessary to make way for new approaches to language that allow for a better understanding of linguistic and cultural diversity and of the communicative functions of English in its entanglements with other semiotic resources in the expanding circle and beyond. Especially since global Englishes can be found in various communicative formats and contexts around the world—including on public signage, in the media, in classrooms, or at the workplace—we need to broaden our conceptualisation of language to allow for a deeper insight into social practices of meaning-making at large.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SS was the sole author of this article. Data used in this article was compiled by SS.

REFERENCES

Adler, M. (2004). *Form und Häufigkeit der Verwendung von Anglizismen in deutschen und schwedischen Massenmedien.* [PhD dissertation]. Jena: Friedrich Schiller University Jena.

Andersen, G. (2015). "Pseudo-borrowings as cases of pragmatic borrowing: Focus on anglicisms in Norwegian," in *Pseudo-English: Studies on False Anglicisms in Europe*, eds C. Furiassi and H. Gottlieb (Berlin: De Gruyter Mouton). doi: 10.1515/9781614514688.123

Baker, W., and Sangiamchit, C. (2019). Transcultural communication: Language, communication and culture through English as a lingua franca in a social network community. *Lang. Intercult. Commun.* 19, 471–487. doi: 10.1080/14708477.2019.1606230

Bateman, J. A. (2008). *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents.* Houndmills: Palgrave Macmillan. doi: 10.1057/9780230582323

Blommaert, J. (2010). *The Sociolinguistics of Globalization.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511845307

Burn, A. (2014). "The kineikonic mode: Towards a multimodal approach to moving-image media," in *The Routledge Handbook of Multimodal Analysis*, 2nd edition, ed C. Jewitt (London: Routledge).

Canagarajah, S. (2013). *Translingual Practice: Global Englishes and Cosmopolitan Relations.* Abingdon: Routledge. doi: 10.4324/9780203073889

Canagarajah, S. (2018a). The unit and focus of analysis in lingua franca English interactions: In search of a method. *Int. J. Bilingual Educ. Bilingualism* 21, 805–824. doi: 10.1080/13670050.2018.1474850

Canagarajah, S. (2018b). Translingual practice as spatial repertoires: Expanding the paradigm beyond structuralist orientations. *Appl. Linguist.* 39, 31–54. doi: 10.1093/applin/amx041

Creese, A., and Blackledge, A. (2011). Separate and flexible bilingualism in complementary schools: Multiple language practices in interrelationship. *J. Pragmatics* 43, 1196–1208. doi: 10.1016/j.pragma.2010.10.006

Crisell, A. (1986). *Understanding Radio.* London: Methuen.

Furiassi, C. (2010). *False Anglicisms in Italian.* Monza: Polimetrica.

Glahn, R. (2002). *Der Einfluß des Englischen auf gesprochene deutsche Gegenwartssprache: Eine Analyse öffentlich gesprochener Sprache am Beispiel von "Fernsehdeutsch".* 2nd edition. Frankfurt am Main: Peter Lang.

Gottlieb, H. (2015). "Danish pseudo-anglicisms: A corpus-based analysis," in *Pseudo-English: Studies on False Anglicisms in Europe*, eds C. Furiassi and H. Gottlieb (Berlin: De Gruyter Mouton). doi: 10.1515/9781614514688.61

Kachru, B. B. (1985). "Standards, codification and sociolinguistic realism: The English language in the Outer Circle", in *English in the World: Teaching and Learning the Language and Literatures*, eds R. Quirk and H.G. Widdowson (Cambridge: Cambridge University Press).

Kress, G. R. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication.* London: Routledge.

Lee, J. S. (2014). English on Korean television. *World Englishes* 33, 33–49. doi: 10.1111/weng.12052

Li, W. (2018). Translanguaging as a practical theory of language. *Appl. Linguistics* 39, 9–30. doi: 10.1093/applin/amx039

Mair, C. (2013). The world system of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34, 253–278. doi: 10.1075/eww.34.3.01mai

Makalela, L. (2013). Black South African English on the radio. *World Englishes* 32, 93–107. doi: 10.1111/weng.12007

Mesthrie, R. (2002). Mock languages and symbolic power: The South African radio series Applesammy and Naidoo. *World Englishes* 21, 99–112. doi: 10.1111/1467-971X.00234

Miller, B. M. (2018). "The pictures are better on radio": A visual analysis of American radio drama from the 1920s to the 1950s. *Histor. J. Film Radio Television* 38, 322–342. doi: 10.1080/01439685.2017.1332189

Onysko, A. (2007). *Anglicisms in German: Borrowing, Lexical Productivity, and Written Codeswitching*. Berlin: De Gruyter. doi: 10.1515/9783110912173

Onysko, A. (2016). Modeling World Englishes from the perspective of language contact. *World Englishes* 35, 196–220. doi: 10.1111/weng.12191

Otheguy, R., García, O., and Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Appl. Linguistics Rev.* 6, 281–307. doi: 10.1515/applirev-2015-0014

Pennycook, A. (2007). *Global Englishes and Transcultural Flows*. Abingdon: Routledge. doi: 10.4324/9780203088807

Pennycook, A. (2017). Translanguaging and semiotic assemblages. *Int. J. Multilingual.* 14, 269–282. doi: 10.1080/14790718.2017.1315810

Pennycook, A. (2020). Translingual entanglements of English. *World Englishes* 39, 222–235. doi: 10.1111/weng.12456

Picone, M. D. (1996). *Anglicisms, Neologisms and Dynamic French*. Amsterdam: John Benjamins. doi: 10.1075/lis.18

Pulcini, V., Furiassi, C., and Rodríguez González, F. (2012). "The lexical influence of English on European languages: From words to phraseology," in *The Anglicization of European Lexis*, eds C. Furiassi, V. Pulcini and F. Rodríguez González (Amsterdam: John Benjamins), 1–26. doi: 10.1075/z.174.03pul

Schaefer, S. J. (2019). Anglicisms in German media: Exploring catachrestic and non-catachrestic innovations in radio station imaging. *Lingua* 221, 72–88. doi: 10.1016/j.lingua.2019.01.002

Schaefer, S. J. (2021a). English on air: Novel anglicisms in German radio language. *Open Linguistics* 7, 569–593. doi: 10.1515/opli-2020-0174

Schaefer, S. J. (2021b). Hybridization or what? A question of linguistic and cultural change in Germany. *Globalizations* 18, 667–682. doi: 10.1080/14747731.2020.1832839

Schneider, E. W. (2007). *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511618901

Schneider, E. W. (2014). New reflections on the evolutionary dynamics of World Englishes. *World Englishes* 33, 9–32. doi: 10.1111/weng.12069

Sharma, B. K. (2012). Beyond social networking: Performing global Englishes in Facebook by college youth in Nepal. *J. Sociolinguist.* 16, 483–509. doi: 10.1111/j.1467-9841.2012.00544.x

Shingler, M., and Wieringa, C. (1998). *On Air: Methods and Meanings of Radio*. London: Arnold.

Stöckl, H. (2016). "Multimodalität - Semiotische und textlinguistische Grundlagen," in *Handbuch Sprache im multimodalen Kontext*, eds N. Klug and H. Stöckl (Berlin: De Gruyter). doi: 10.1515/9783110296099-002

Sultana, S., and Dovchin, S. (2017). Popular culture in transglossic language practices of young adults. *Int. Multilingual Res. J.* 11, 67–85. doi: 10.1080/19313152.2016.1208633

Van Leeuwen, T. (1999). *Speech, Music, Sound*. London: Macmillan Education. doi: 10.1007/978-1-349-27700-1

Yang, W. (1990). *Anglizismen im Deutschen: Am Beispiel des Nachrichtenmagazins Der Spiegel*. Tübingen: Niemeyer. doi: 10.1515/9783111676159

Zenner, E., Speelman, D., and Geeraerts, D. (2012). Cognitive sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognit. Linguistics* 23, 749–792. doi: 10.1515/cog-2012-0023

Check for updates

Perception and Reinterpretation of English Song Lyrics by Native Speakers of Japanese: A Case Study of Samples From the TV-Show *Soramimi-Hour*

Johannes Scherling[1]*, Lisa Kornder[1] and Niamh Kelly[2]

[1] *Department of English Studies, University of Graz, Graz, Austria,* [2] *Department of Languages and Linguistics, University of Texas at El Paso, El Paso, TX, United States*

Cross-linguistic mondegreens occur when foreign song lyrics are misperceived and reinterpreted in the listener's native language. In Japan, such humorous reinterpretations of non-native song lyrics are known as *soramimi* (空耳, "mishearing"). Word plays of this kind do not only have an entertaining character for listeners, but they also offer a valuable source to identify and describe potential phonological processes which can be observed in native Japanese listeners' adaptations of English song lyrics into Japanese. We hypothesized that the reinterpretation of English song lyrics by listeners is a result of the perception of non-native auditory input through the first language's phonological and morphological system. That is, misperceptions do not occur arbitrarily, but are governed by the phonological and morpho-phonological rules of the listener's first language system. To test this hypothesis, we examined a corpus containing 60 English-Japanese mondegreens taken from the Japanese TV-show *Soramimi Awā* (*Soramimi Hour*). Results confirmed our hypothesis: The Japanese adaptations were observed to follow different phonological processes which aimed to subject the non-native auditory input to the phonological rules of Japanese.

Keywords: Japanese, English, cross-linguistic mondegreens, phonological processes, Japanese phonotactics, top-down processing, language contact

INTRODUCTION

In the present paper, we aimed to find out which phonological processes underlie native Japanese listeners' misperceptions and reinterpretations of English song lyrics, known as *soramimi* in Japanese (空耳, "mishearing"). This investigation is based on the hypothesis that native Japanese listeners reinterpret English auditory input through the phonological filter of their native language. In this context, we argue that—similar to the adaptation of English loanwords into Japanese—the misperception of foreign language lyrics is governed by the phonological rules of listeners' first language system. A corpus of 60 English-Japanese *soramimi* was analyzed and, based on a comparison of the English original and the Japanese misperceived lyrics, categorized according to one of four (morpho-)phonological processes, that is, sound substitution, vowel insertion, segment deletion, and boundary transgression. After briefly outlining the history of language

contact between Japanese and English, aspects that relate to Japanese phonotactics will be discussed. In Section Integration Processes for Foreign Language Material in Japanese, integration processes which are relevant when it comes to adapting foreign language material into Japanese are described. This is followed by an outline of English phonotactics (Section English Phonotactics) and briefly comparing the English and Japanese vowel and consonant inventories (Section The English and Japanese Sound Inventories). Section Mondegreens, *Soramimi*, and Speech Perception explains what mondegreens and *soramimi* are and how they are related to speech perception. In the subsequent section, we outline our methodology and provide an analysis and discussion of our findings.

A SHORT HISTORY OF LANGUAGE CONTACT BETWEEN JAPANESE AND ENGLISH

Though language contact in Japan has been a comparatively rare occurrence due to its geography as an island nation, when it did happen, it affected the Japanese language considerably. The best example for this is the intensive contact between the Japanese and the Chinese language that took place from the first century AD (Hoffer, 2002, p. 29; Loveday, 1996, p. 264-65). This profound cultural exchange not only led to the introduction of Buddhism by the sixth century—mostly *via* Sino-Korean influence (Loveday, 1996, p. 39-40)—but also resulted in the wholesale adoption of the Chinese writing system and lexicon [∼50% of the contemporary Japanese lexicon are of Chinese origin (Morrow, 1987; Stanlaw, 2004; Scherling, 2012)], which over time the Japanese "were able to develop and rework […] according to their own needs" (Loveday, 1996). Out of these Chinese characters, Japanese then developed its two syllabaries, namely Hiragana and Katakana (Matsuda, 1985; Stanlaw, 2004), in order to enable people to read texts without having to be able to read the complex Chinese characters only known to the learned at that time. These syllabaries represented all possible sound combinations known to Japanese. Applying the phonotactic constraints determined by the syllabaries to foreign words then became the blueprint for assimilative processes that—alongside morphological processes such as clipping and blending, based on the Chinese model, where loanwords are treated as uninflected and bound bases (Loveday, 1996, p. 138) – have shaped loanword integration[1] in Japan to this day (Stanlaw, 2004).

Contact to European languages came rather late. Only when Europe's empires became seaborne did a chance for language contact arise. The first to reach the shores of Japan—in the sixteenth and seventeenth century—were the Portuguese, followed by the Spanish and the Dutch, all of which left behind a number of loanwords still in use today, such as "pan" (パン, Portuguese for "bread"), "kasutera" (カステラ, Spanish for

"sponge cake") or "garasu" (ガラス, Dutch for "glass"). Japan then delved into a self-imposed period of isolation, a result of arguably well-founded fears of meddling or even colonization (Stanlaw, 2004, p. 46-47). It took Japan some 250 years before opening up to the world again—under the threat of force. In 1853, US-American Commodore Matthew Perry sailed into Edo Bay (what is nowadays Yokohama Bay) with heavily armed gunships to successfully force Japan to open its ports to international trade again (LaFeber, 1997, p. 13). It was through this traumatic experience that the Japanese government realized its lag in technological advancement to the world and felt the need to "modernize" and "westernize" by adopting and assimilating all things Western (Loveday, 1996, p. 62; Stanlaw, 2004, p. 52-56). Increasingly, then, by the early twentieth century, the means of "modernization" became the English language, with words like "takushi-" (タクシー, Eng. "taxi"), "rajio", (ラジオ, Eng. "radio") or "jiruba" (ジルバ, Eng. "jitterbug") entering the language in quick succession (Stanlaw, 2004). After a short, but all the more intense intermezzo of hostilities during the 1930s and 1940s which culminated in the Pacific War, during which English loanwords were partially replaced by words of Chinese and Japanese origin (Oishi, 1992), and after Japan's defeat, epitomized through the devastation of two atomic bombs, the English language celebrated a phenomenal comeback; this time as the perceived key to attaining the much admired living standards of the "American Way of Life" (Loveday, 1996; Dower, 1999). In the aftermath of the war, the number of English-based loanwords started to increase steeply, reaching proportions of more than 10% of the entire Japanese lexicon (Scherling, 2012), a number that might be even higher if the many hybrid words (such as "shirobai" (白バイ), meaning "motorcycle policeman", a combination of "shiroi"—白い, "white"—and "bike") that contain English language material are taken into account.

One consequence of this modernization—and a gradual but incremental effect leading back as far as the Meiji Restoration of 1868 (Pintér, 2015, p. 175)—has been innovations in the Japanese syllabary system. These became necessary in order to take into account as of yet unknown sounds and sound sequences that were needed to assimilate foreign words into the Japanese linguistic matrix, such as "fi" (フィ), as in "fire" (フィレ, Eng. "filet") which had previously been rendered as "hire" (ヒレ), or "va" (ヴァ), as in "vaiorin" (ヴァイオリン, Eng. "violin") which used to be "baiorin" (バイオリン) (Stanlaw, 2002). These innovations have undoubtedly made it easier to adapt English words into Japanese phonology. They have, however, changed little regarding the very basic phonological differences between the two languages, which instantly necessitate adaptations in the sound structure of English-based loanwords on their arrival in Japanese and which will be the focus of the following section.

JAPANESE PHONOTACTICS

The Japanese language is—with the exception of some more differentiated views (Warner and Arai, 2001)—widely regarded

---

[1] From a phonetic/phonological perspective, loanword adaption can be described as "a process that applies during speech perception and that maps non-native sound structures onto the phonetically closest native ones" (Peperkamp, 2005, p. 2).

a mora-timed language "where each mora is supposed to take an equal duration of time" (Kubozono, 2002, p. 33). A mora is a phonological unit describing syllable weight (Hogg, 1992) and is also called a phonemic syllable (e.g., Pike, 1947; Kubozono, 2002). It differs from the syllable in that, while every syllable consists of (at least) one mora, not every mora can constitute a syllable (Kubozono, 2015a, p. 63). For instance, while "Nagasaki" consists both of 4 syllables and 4 morae, a word like "Tōkyō" has only 2 syllables but 4 morae, due to the two long vowels (essentially To-o-kyo-o). For the present investigation, understanding the importance of morae is vital as it "plays a crucial role in speech perception" (Kubozono, 2002, p. 39) and because Japanese speakers have been shown to "respond to mora-sized units more readily than syllable-sized units" (Kubozono, 2002 citing Hayashi and Kakehi, 1990).

Owing to its syllabic structure (reflected in its writing system which requires consonants to be represented along with a vowel and not individually, while vowels can be represented individually), the phonotactic constraints that apply to Japanese words or words that are adopted into Japanese are rather rigid. Japanese permits only open syllables, with the sole exception of the nasal /N/. This means that the only permissible sound sequences are (C)V, (C)VV, and (C)VN (Dupoux et al., 1999, p. 2), with the caveat that not all consonants can combine with all vowels (Pintér, 2015, p. 174). For example, the consonant /t/ can combine with /a/, /e/ and /o/, but not with /i/ and /u/, whereas /s/ cannot combine with /i/, and /w/ can only combine with /a/. Despite the innovative changes to the syllabary in recent decades (see above), this constrains and influences the integration of non-Japanese words into the writing system, in particular since, as Stanlaw suggests, there is a recent trend to use the more conservative spelling, and the new syllabary characters are not all that widespread (Stanlaw, 2004, p. 95 citing Inoue, 1996, p. 196-97).

Therefore, when novel words enter the Japanese language from another language, the first impact they experience is on their phonology, as many sound combinations from languages such as German or English are not permissible in Japanese and can therefore not be represented without some changes. Since Japanese, as outlined above, disallows consonant clusters or closed syllables, such consonant clusters need to be broken up or complemented with vowels so that they can fit Japanese phonological/phonotactic structure. Another theory maintained by Dupoux et al. is that, in fact, Japanese phonotactic constraints lead Japanese speakers to "perceive epenthetic [u] vowels within consonant clusters" (Dupoux et al., 1999, p. 11) to adapt foreign language input to their language's phonotactics. The exact phonological processes which are relevant in this context will be described in more detail below.

Another important aspect of Japanese phonology pertains to the production of vowels in certain environments, a phenomenon that is called "devoicing". Devoicing means that vowels that are typically pronounced with the vocal folds vibrating under certain conditions lose that voicing, one of those conditions being speaking rate (Fujimoto, 2015, p. 215). According to Fujimoto, devoicing affects the high vowels /i/

and /u/ in the morae /pi/, /pu/, /ki/, /ku/, /ɕi/,[2] /ɕu/, /tɕi/, /tɕu/, /hi/, /su/, /tsu/, and /hu/, in particular when they are followed by a voiceless consonant (ibid). Hence, a word such as "shita" (下, "below") would tend to be pronounced like [ɕita], or "shuto" (首都, "capital") like [ɕu̥to], to the extent that the vowel is barely perceivable anymore. Devoicing also applies in word-final environments, such as "desu" (です, "to be"), which is pronounced like [desɯ̥]. Acoustic examinations, however, confirm that, despite the devoicing and possible subsequent reduction in distinction between /i/ and /u/, native Japanese speakers are both able to differentiate between and identify different vowels with identical onset consonants (e.g., /ɕi/ and /ɕɯ/) (Fujimoto, 2015, p. 194). In this respect, what is of interest in the context of the present study is whether native Japanese listeners will show a tendency to "hear" devoiced vowels in consonant clusters from English words in their "misheard" song lyrics as well, seeing that speaking rate in songs tends to be faster the faster the melody is (Jungers et al., 2002) and given that "faster speech facilitates devoicing in non-general conditions and atypical consonantal conditions" (Fujimoto, 2015, p. 206).

INTEGRATION PROCESSES FOR FOREIGN LANGUAGE MATERIAL IN JAPANESE

As discussed above, the particular phonotactic rules and constraints of the Japanese language necessitate fundamental changes to the phonological structure of loanwords. This is a result of its syllabary alphabet which cannot, for example, represent individual consonants or consonant clusters. According to Loveday (1996, p. 114), on arrival in Japanese, "the syllabic structure of English is radically altered and new sets of morpho-phonemic and phonotactic patterns are introduced that are not always regular or predictable". In the following, the most frequent phonological processes that such loanwords undergo when adopted will be discussed with a view to misprisions of English song lyrics, which we argue are subject to similar strategies and processes.

The two most frequent processes affecting loanword integration are sound substitutions and vowel epenthesis, or insertion (Loveday, 1996; Nian and Jubilado, 2011; Scherling, 2013). Sound deletion also plays a certain role, in particular for unstressed syllables or in environments where a consonant coda is not perceived or not realized (Shoji and Shoji, 2014). Vowel epenthesis, as argued above, is a natural consequence of Japanese orthography, phonotactics and language perception. When disallowed consonant clusters or consonant codas are encountered, these are broken up by inserting vowels, mostly /i/, /o/ and /u/, to ensure open syllables. According to Stanlaw (2004, p. 74), the choice of the epenthetic vowel is a function

---

[2]The source referred to here, but also others we cite in this paper, use the IPA symbols /ʃ/, /ʒ/, /tʃ/ and /dʒ/ to refer to Japanese alveolo-palatal fricatives and affricates. As Kubozono (2015a, p. 6) states, this is debated but not uncommon, yet for the sake of consistency with our own transcriptions of the respective consonants (/ɕ/, /ʑ/, /tɕ/, /dʑ/) we have decided in favor of replacing the IPA symbols used in the original transcriptions with the ones we employ in ours.

of the phonological environment: /u/, called the "context-free default epenthetic vowel" by Shoji and Shoji (2014, p. 3), is inserted after most consonants. While /o/ usually follows after the alveolar consonants /t/ and /d/, /i/ is added after affricates /tɕ/ and /dʑ/—the two being "context-dependent epenthetic vowels" (Shoji and Shoji, 2014, p. 3). Therefore, an English word such as "infrastructure" would be rendered as "infurasutorakucha-" (インフラストラクチャー) in Japanese, a word like "bridge" would become "burijji" (ブリッジ, where the duplicated "j" is a geminated voiced affricate), and "drugstore" would become "doragusutoa" (ドラグストア). Clearly, such inserted vowels make the loanword longer in terms of syllables and also incrementally remove it phonologically from the source word, so that they would become difficult to identify as English words for native English speakers. However, as Dupoux et al. (1999) have argued, for Japanese speakers, such vowels appear to be actually perceived, which may help explain some of the mondegreens to be discussed later.

The second process, substitution, applies to vowels and consonants that are non-phonemic in Japanese and therefore need to be replaced with phonologically similar, approximate sounds available in the Japanese sound inventory. There is a considerable number of sounds that are unknown in Japanese, such as the interdental fricatives /θ/ and /ð/, the lateral approximants /r/ and /l/, and the labiodental fricative /v/. There are also illicit syllables like /tɪ/ or /dɪ/, as well as non-existent vowels, such as /ɛ/ or /æ/ (Matsuda, 1986, p. 49; Stanlaw, 2004, p. 74). Foreign words that contain such sounds are subject to sound substitution. The choice of the sound that will substitute for the non-phonemic one is variable and considerably depends on the phonological environment. As previously mentioned, not all consonants combine with all vowels in Japanese. Hence, the interdental /θ/ can be substituted by sounds as different as /s/ (third > sādo), /ɕ/ (thick > shikku) or /ts/ (thulium > tsuriumu) (Stanlaw, 2004, p. 74). This is because the combination /sɪ/ is disallowed in Japanese, and must be realized as /ɕɪ/. For "thulium" (see above), the reason is possibly that the word was imported from German rather than English and hence the /tu/ sound sequence had to be replaced by /tsu/ as it is not a permitted sound combination in Japanese. For vowels, the patterns are more regular, but somewhat dependent on whether the perception of the foreign word happens visually or aurally[3]. An example of that is "Christmas" (/krɪsməs/), in which case the word was likely first encountered visually, since the hardly perceptible mid central vowel /ə/ is rendered as /a/ in Japanese (/kɯrisɯmasɯ/). Nian and Jubilado (2011, p. 101) and Tsuchida

(1995, p. 147-48), for instance, outline the following broad regularities when it comes to the substitution of vowels:

Eng. /ɜ/ > Jpn. /a/: e.g., "bird" /bɜːd/ > bādo /baːdo/
Eng. /æ/ > Jpn. /a/: e.g., "taxi" /tæksi/ > takushi- /takɯɕiː/
Eng. /ɛ/ > Jpn. /e/: e.g., "pet" /pɛt/ > petto /petto/
Eng. /ɔ/ > Jpn. /o/: e.g., "boarding" /bɔrdɪŋ/ > bodingu /boːdiŋgɯ/

Similarly, English diphthongs—considering that most scholars agree that diphthongs do not exist in Japanese[4]—are also replaced, for example by long monophthongs, such as "cake" (/keɪk/) becoming "ke-ki" (ケーキ, /keːki/) or "shake" (/ʃeɪk/), realized as "she-ku" (シェーク, /ɕeːkɯ/) (Shoji and Shoji, 2014, p. 6). As with consonants, however, there is also a certain flexibility with how vowels are represented in Japanese, largely dependent on the English variant that the word in question was provided by. Hence, in words such as "ballet" (AE /bæleɪ/) and "volleyball" (AE /vɑlibɔl/), the different vowels in English will be represented by the same vowel (/a/) in Japanese: both are rendered as "bare-" (バレー, /bareː/), with the two different consonants /b/ and /v/ being collapsed into one (/b/). Clearly, substitution and vowel epenthesis are closely interlinked, as the choice of the substituting sound may determine the nature of the epenthetic vowel based on the phonotactic constraints of the language. These underlying rules, in turn, may then affect the perception of the sound structure of non-Japanese words.

The last phonological process to be discussed here is sound deletion. As Shoji and Shoji (2014, p. 3) state, Japanese loanwords "prefer epenthesis to deletion," yet deletion is known to occur, for example, with inflectional suffixes in English, such as "smoked salmon", which becomes Japanese "sumo-ku sa-mon" (スモークサーモン), or similarly "corned beef", which is Japanese "ko-n bi-fu" (コーンビーフ) (Shoji and Shoji, 2014, p. 9). The same applies to English "skiing" which is realized in Japanese as "suki-" (スキー), "salaried man" which becomes "sarari-man" (サラリーマン), or the phrase "three strikes" which is Japanese "suri- sutoraiku" (スリーストライク) (Stanlaw, 2004, p. 75). In other cases, final consonants may become the target of deletion, such as in the Japanese renderings of "alright" as "o-rai" (オーライ) or "don't mind" as "donmai" (ドンマイ), as well as "handkerchief" as "hankachi" (ハンカチ). Shoji and Shoji (2014, p. 9) argue that such deletions may be due to a lack of perception of the consonant coda of the closed syllable. Here, an argument can be made that these phonological processes, being related specifically to the mental perception of foreign words through the filter of Japanese phonotactics, apply not only to loanword integration, but more broadly to foreign language perception in general, and may thus also apply to the perception of song lyrics in languages other than Japanese. They further suggest that the choice between epenthesis and deletion often depends on whether the language input occurred *via* the aural or the visual channel, an argument that Stanlaw also proposes when he distinguishes between borrowing by eye

---

[3]Previous research suggests that orthography plays a crucial role in loanword adaptations from one language to another (e.g., Kaneko, 2006; Venedlin and Peperkamp, 2006). Kaneko (2006), for instance, investigated the role of orthography in the selection of Japanese vowels in English loanwords and found that presenting phonological forms together with their spelling indeed had an impact on which vowels native Japanese listeners were likely to select. For example, in an oral condition (i.e., presenting phonological forms without spelling), listeners adapted the English low back vowel [ɑ] as the phonetically closest Japanese targets [a] or [aː]. However, if presented in a mixed condition (i.e., phonological forms with spelling), listeners were more likely to adapt the English target vowel as Japanese [o] due to orthographic influences.

[4]Some scholars argue that there are vowel sequences in modern Japanese which are acoustically similar to diphthongs; this particularly applies to the vowel sequences /ai/, /oi/, and /ui/ (see e.g., Kubozono, 2005).

and borrowing by ear (Stanlaw, 2004, p. 91-2). For example, the English word "pudding" exists in Japanese both as a borrowing by eye ("pudingu", プディング) as well as a borrowing by ear ("purin", プリン), the second of which shows a case of consonant deletion. Similarly, this applies to loanwords such as "pokke" (ポッケ, Eng. "pocket"), where the /t/ was deleted, or "hankachi" (ハンカチ, Eng. "handkerchief"), where word-final /f/ was omitted (Shoji and Shoji, 2014, p. 9). Hence, aural perception—as is the case in *soramimi*—appears to make deletion more frequent than it is when words are encountered *via* the written mode.

We wish to raise one last issue here, which we are proposing may happen in foreign language perception. It connects the process of deletion of final consonants in the perception of foreign words with a morphological issue that may arise in the perception of longer and uninterrupted stretches of spoken language, such as in songs, and which is referred to as boundary transgressions. Final consonant deletion is possible in isolated, individual words because the consonant coda is not perceived when the word is pronounced out of context with other words. In connected speech, however, when the word in question is followed by another word, the final consonant may well be perceived as being the onset of the following word, in particular when this onset is realized by a vowel. This means that if an English word ends in a consonant coda, which may well be deleted if perceived in isolation, this consonant might—instead of being omitted—be reanalyzed as the onset of the following word, thus potentially leading to a mishearing. An example for this taken from our corpus would be the perception of the English phrase "Nothin' can hold us and" as "Naze ka hon dashita" (meaning "For some reason, I published a book"), where the final consonant of "hold" was reanalyzed as the initial consonant of the following word "dashita".

ENGLISH PHONOTACTICS

The phonotactic patterns of English allow complex syllable structures involving consonant clusters. Consonant clusters are licit in onsets as well as in codas, that is, English syllables can be open or closed. Onset clusters generally have a rise in sonority toward the syllable nucleus (although clusters such as /st/ are exceptions to this generalization), while coda clusters have a fall in sonority from the nucleus (e.g., Zec, 2007). As such, English allows more complex syllables than Japanese, meaning that if English sequences are to be interpreted as Japanese words, first language (L1) Japanese speakers have to adapt them to the stricter syllable structure of Japanese. Further, English is generally considered a stress-timed language (e.g., Pike, 1945; Bolinger, 1965) (although dialectal variation occurs, for example, with Jamaican English being described as more syllable-timed Wassink, 2001). It should be noted that the division of languages into stress-timed, syllable-timed and mora-timed has been questioned (e.g., Arvaniti, 2009) and there is no clear consensus on what phonetic measurements can clearly represent these proposed distinctions. However, this topic continues to be discussed and investigated from the perspectives of both

production and perception, taking into account such variables as speaking rate, syllable structure and phrasal prominence (e.g., Turk and Shattuck-Hufnagel, 2013; Fuchs, 2016). Assuming the difference in rhythm between English and Japanese is valid, speakers of a mora-timed language such as Japanese may parse the syllables of English differently from how English speakers would.

THE ENGLISH AND JAPANESE SOUND INVENTORIES

The General American English vowel inventory comprises 14 to 15 distinctive vowels, that is, /iː/, /ɪ/, /ɛ/, /æː/, /ɑː/, /ʌ/, /ɔː/, /ʊ/, and /uː/, including the rhotic vowel /ɝ/, the mid-central vowel /ə/, and the diphthongs /eɪ/, /oʊ/, /aɪ/, /aʊ/, and /ɔɪ/ (Ladefoged, 2005). It should be noted though that in many American English varieties (e.g., Californian English, see Hagiwara, 1997) the distinction between /ɑː/ and /ɔː/ is neutralized in the direction of [ɒː] (Nishi et al., 2008). In contrast to American English, the modern Tokyo Japanese vowel inventory includes five distinctive monophthongal vowels only, that is /i/, /e/, /a/, /o/, and /ɯ/ (e.g., Ladefoged and Johnson, 2014; Kubozono, 2015b; see **Figure 1**), which are differentiated according to vowel height (high, mid, low) and tongue backness (front vs. back). Corresponding to the five short monophthongs, Japanese has five long monophthongs, i.e., /iː/, /eː/, /aː/, /oː/, and /ɯː/ (Kubozono, 2015b, p. 3).

Table 1 displays the Japanese and American English consonant inventory, respectively. Both inventories share the plosive consonants /p b t d k g/, the nasals /m n ŋ/, the approximants /j/ and /w/, as well as the alveolar fricatives /s/ and /z/. The English inventory has additional fricative consonants which do not exist in Japanese, namely /f v θ ð ʃ ʒ/. Japanese, by contrast, includes the bilabial fricative /Φ/, as in "fune" (船, Eng. "boat"), and the alveolo-palatal fricatives /ɕ/ and /ʑ/, as in "shimbun" (新聞, Eng. "newspaper") and "jagaimo" (じゃがいも, Eng. "potato"), respectively. Unlike English, Japanese does not have a contrast between /l/ and /r/, and there is no consonant in Japanese which exactly corresponds to either English /l/ or English /r/ (see Guion et al., 2000; Ohata, 2004). According to Guion et al. (2000, p. 2711), phonologically, the Japanese liquid consonant /r/ "might be considered similar to both English /l/ and /ɹ/ [sic!]", but is typically produced as an alveolar or postalveolar apical tap [ɾ] (e.g., Vance, 1987; Labrune, 2012).

Unlike English which has two affricates only, i.e., /tʃ/ and /dʒ/, both produced at a postalveolar place of articulation, the Japanese inventory comprises four affricate consonants, namely the voiceless and voiced alveolar affricates /ts/ ("tsume", Eng. "nail") and /dz/ ("zurui", Eng. "sly"), and the alveolo-palatal affricates /tɕ/ ("chuugoku", Eng. "China") and /dʑ/ ("jikan", Eng. "time").

While the English nasal inventory includes two consonants only, i.e., /n/ and /ŋ/, Japanese has a larger nasal inventory resulting from allophonic variation of the moraic nasal /N/, which does not have a specified place of articulation (e.g., Otake et al., 1996; Labrune, 2012; Kubozono, 2015a). For instance, /N/ is typically realized as the uvular consonant

FIGURE 1 | American English (green; adapted from Ladefoged and Johnson, 2014) and Japanese (blue; adapted from Kubozono, 2015b) vowel inventories. Vowels shared by both inventories are marked in gray.
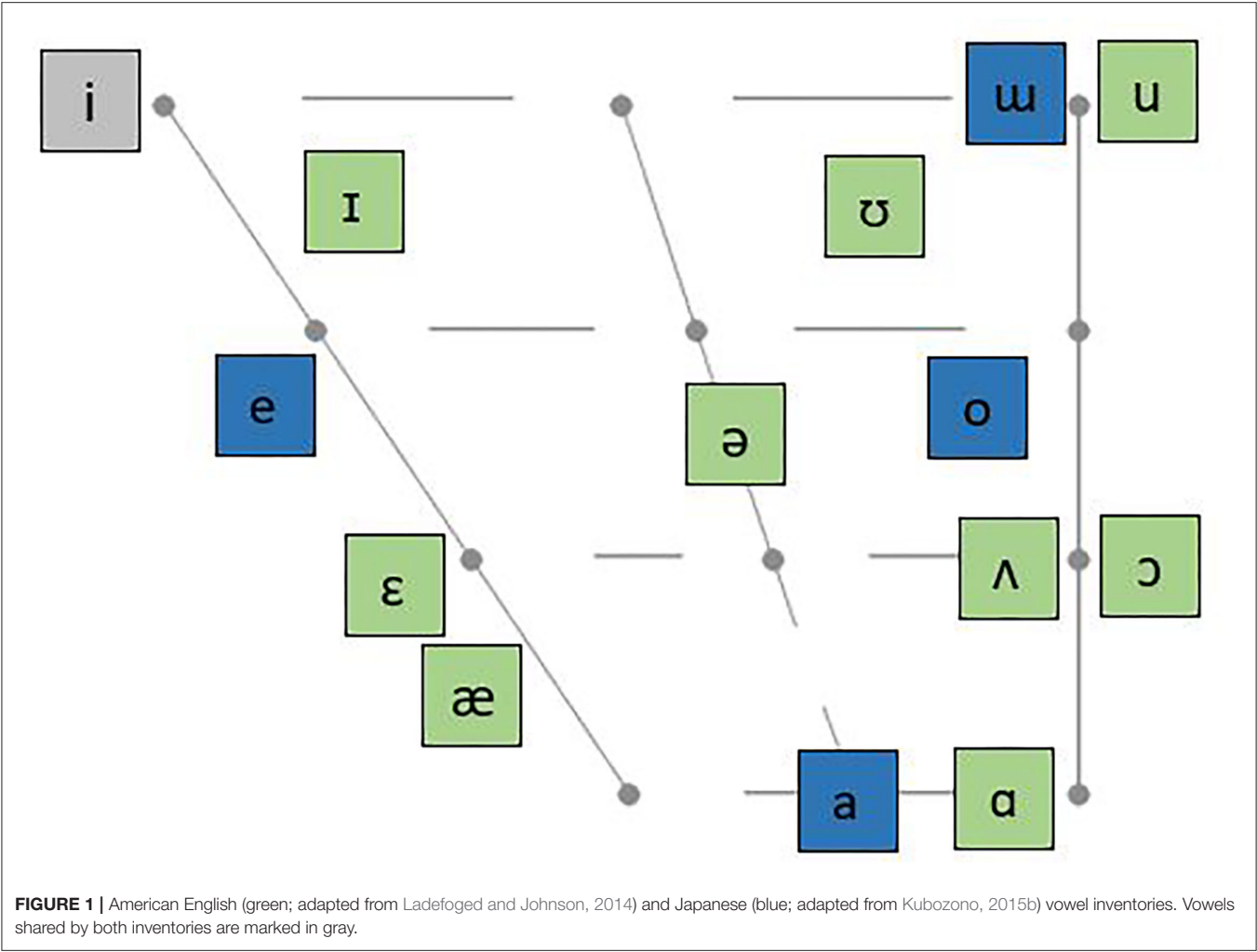
TABLE 1 | English (green) and Japanese (blue) consonant inventories.

	Bilabial		Labiodental		Dental		Alveolar		Alveolo-palatal		Postalveolar		Palatal	Velar		Uvular	Glottal
Plosive	p	b	t	d										k	g		ʔ
Nasal		m					n								ŋ	N	
Affricate							ts	dz	tɕ	dʑ	tʃ	dʒ					
Tap or Flap							ɾ										
Fricative	ɸ		f	v	θ	ð	s	z	ɕ	ʑ	ʃ	ʒ					h
Approximant							ɹ						j				
Lateral approximant							l							w			
Moraic consonants									/Q/, /N/								

Consonants shared by both inventories are marked in gray.

[N] particularly in slow speech. When occurring before the stops /p b m/, /N/ is realized as the bilabial nasal [m], when preceding the alveolar consonants /t d n/, the moraic nasal is produced as [n], and when occurring before the velar plosives /k g/, it is realized as [ŋ] (see Labrune, 2012, p. 133-4, for a detailed overview).

MONDEGREENS, *SORAMIMI,* AND SPEECH PERCEPTION

Mondegreens are often described as phenomena of auditory illusion (Kentner, 2015), and refer to listeners' misperceptions of words or phrases in songs presented in the listeners' native

language ("within-language misperceptions", Beck et al., 2014, p. 2). However, misperceptions of song lyrics are not restricted to songs in the listener's native language, but might also occur across different languages, which Kentner (2015, p. 1) refers to as "cross-linguistic mondegreens". This means that listeners may perceive foreign lyrics, including both individual words and complete phrases (Otake, 2007), through their native language and adapt the foreign stimuli to their native language system. Typically, "mondegreening", resulting from either within-language or across-language misperceptions, leads to the creation of new, usually amusing, meanings (e.g., Beck et al., 2014). The term "Mondegreen" goes back to Wright (1954) who described her mishearing of the original line "layd him on the green" in the seventeenth-century Scottish folk song "The Bonny Early O' Murray" (see Olson, 1997) as "Lady Mondegreen". In Japanese, mondegreens are known as *soramimi* (空耳, "mishearing"), which has lent its name to the popular Japanese TV show *Soramimi Hour* (*soramimi awā*). For the sake of entertainment, Japanese listeners are invited to submit misheard English song lyrics to the show which are then presented in short comical video clips to the audience and which are rated and awarded by the show hosts based on the degree to which they share the mishearing.

From a linguistic point of view, mondegreens and *soramimi* can be considered a "valuable tool to induce plasticity within the auditory system" (Beck et al., 2014, p. 2), but only few studies so far have systematically examined the actual processes underlying the phenomenon of misperceiving song lyrics and adapting foreign lyrics to the listener's native language system. To the best of our knowledge, to date there is only one study (Otake, 2007) that investigated phonological processes involved in listeners' misperceptions of foreign song lyrics: Otake (2007) examined a corpus of 194 English-Japanese *soramimi* containing interlingual near-homophonic words and phrases, i.e., English words and phrases which were misperceived as near-homophonic Japanese targets by Japanese listeners (e.g., the English source lyric "psycho" was perceived as Japanese "saikō" ['saɪkoʊ], "the highest", see Otake, 2007, p. 778). Otake identified different reoccurring phonological processes, including segment omissions, sound substitutions and additions (see Sections Integration Processes for Foreign Language Material in Japanese and Methodology, this paper), underlying the Japanese adaptations of English auditory input identified in the *soramimi* samples. As Otake (2007, p. 779) concludes, the *soramimi* examples revealed "the same kinds of processes as appear in adaptation of foreign loan-words in Japanese" (see also Peperkamp et al., 2008), showing that Japanese listeners may perceive English words and phrases through the phonological filter of their native language.

In order to give further insights into the phenomenon of cross-linguistic auditory misperception and the phonological and morphological processes involved, the present study set out to analyze 60 examples of English-Japanese *soramimi* taken from the above-mentioned Japanese TV show *Soramimi Hour* (see Section Methodology, for a detailed description of the speech corpus).

CUES TO PROCESSING CONNECTED SPEECH

When processing spoken language, listeners mainly use the acoustic signal, but may also use visual cues (McGurk and MacDonald, 1976) and even tactile cues, when available (Gick and Derrick, 2009). The McGurk Effect has been shown to occur among adult listeners, whereby speech processing can be affected by visual cues—in this study, silent videos of speakers producing stop consonants—that do not match the auditory signal (McGurk and MacDonald, 1976). Processing spoken language is a challenge which contends with factors that affect production, including coarticulation, emotional effects, inter-speaker variability, accentedness, speaking rate, and background noise (e.g., Ladefoged and Broadbent, 1957; Mann and Repp, 1980; Brent et al., 2013; Dame et al., 2013). Therefore, listeners use as many cues as possible in order to avoid mishearing and misunderstanding the signal. Using these cues are examples of bottom-up processing, but listeners also use knowledge about the structure and patterns of the language in order to interpret the signal, a process known as top-down processing (e.g., Warren, 1970; Ganong, 1980). For example, if a person listening to English hears "the", they know that the following word is likely to be a noun or an adjective. It has been shown that a sound that is at the acoustic boundary between two phonemes will be interpreted as one phoneme or the other based on whether the resulting word would form a real word (Ganong, 1980). For example, if a sound is at the acoustic boundary between English /d/ and /t/, if it is presented as the onset to "ask", listeners will tend to hear "task" rather than "dask", because only the former is a real word. As such, speech processing is influenced by knowledge of the language.

Especially in a noisy environment where the acoustic signal is degraded, listeners try to use as many cues as possible to interpret speech. Top-down processing, therefore, is likely to be relied on more when the acoustic signal is noisy. As such, due to the combined effect of music in the signal, song lyrics are potential triggers for mishearing lyrics, also known as mondegreens (see Section Mondegreens, *Soramimi*, and Speech Perception). Balmer (2007) proposes the Phonetic Ambiguity Hypothesis which states that the phonetic ambiguity present in song lyrics is what triggers mondegreens. Also, in listening to songs, listeners process the words and melodies interactively (Gordon et al., 2010; Tang, 2015), and further, listeners do not have the contextual cues that would be available in conversational speech (Tang, 2015). Furthermore, due to the less clear signal in music, lyrics are ideal targets for manipulation when paired with visual cues and written words. Research has found that phonological as well as orthographic factors can have a priming effect in picture-word interference tasks (Chéreau et al., 2007; Damian and Bowers, 2009). Such an effect may extend to *soramimi*, which are presented as visual scenes as well as Japanese words highlighting the misheard lyrics. The presentation of these multiple cues is likely to prime the intended interpretation for viewers. It should also be noted that previous research (e.g., Starr and Shih, 2017) has suggested that Japanese speech produced in

TABLE 2 | Example *soramimi* (No. 190), corresponding IPA transcriptions and phonological processes identified in this example.

	Lyrics	IPA	Processes
English original	*I knocked on every door*	[anɑktɑ/nɛ/vɹ̩ido:]	4 SUB 1 INS 1 DEL
Japanese misperception	*anata nee buri dou*	[anata/nee/bɯᵝɾʲi do:]	1 BTR 8 NoCh

Phonological processes (SUB, INS, DEL) are color-coded.

songs may be different from spoken speech with regard to certain aspects such as phonology; the fact that the input analyzed in the present study consists of sung language may therefore also have an impact on native Japanese speakers' perception thereof.

A last aspect to consider when it comes to processing connected speech are potential influences of social information. Strand (1999), for example, found that gender stereotypes maintained by listeners significantly alter their perception along the /s/–/ʃ/-continuum. Similarly, Johnson et al. (1999) observed that speaker gender influenced how listeners located the phoneme boundary between the English vowel targets /u/ and /ʊ/. With regard to the (mis-)perception of song lyrics, as analyzed in the present study, listeners' perceptions of individual sounds and/or words might therefore—at least to some extent—be affected by the gender (and potentially other social information) of the singer.

METHODOLOGY

In order to examine which phonological and morphological processes are involved in the misperception of English song lyrics by Japanese listeners, a speech corpus of 60 English-Japanese cross-linguistic mondegreens[5] (*soramimi*) was analyzed. This corpus was assembled based on a large pool of "felicitous" *soramimi*[6] from various episodes of the Japanese TV-show *Soramimi Hour*, broadcast since 1992, from which 60 examples were then randomly selected. After downloading the samples from the online platform YouTube, both the original English lyrics (source lyrics) and the Japanese misperceptions were transcribed into IPA by two trained phoneticians based on the extracts presented in the clips (see an example in **Table 2**). The individual samples represent short phrases from different English song lyrics and the corresponding Japanese mishearings, which differed in syllable number (English: $M = 6.78$, $SD = 2.39$; Japanese: $M = 7.22$, $SD = 2.55$), depending on where the misperception occurred. Based on a comparison between the English source lyrics and the Japanese misperceptions, the adaptations and changes identified in the Japanese samples were categorized according to one of three phonological processes, i.e.,

sound substitution (SUB), vowel insertion (INS), and segment deletion (DEL) (see Otake, 2007; see **Table 2**). As outlined in Section Integration Processes for Foreign Language Material in Japanese, these processes are most frequently observed when it comes to adapting English loanwords to the Japanese language (e.g., Loveday, 1996; Nian and Jubilado, 2011; Scherling, 2013; Shoji and Shoji, 2014). Sound substitutions are likely to occur when English vowels (e.g., /ɜ/, /æ/) and consonants (e.g., /θ/, /ð/, /r/, /l/, /v/) which are not part of the Japanese inventory are encountered and are replaced by phonologically similar Japanese sounds (see e.g., Stanlaw, 2004; Pintér, 2015). Alongside replacing an English sound with a Japanese target, an epenthetic vowel—typically /i/, /o/, or /u/—might be inserted in an English source word to break up consonant clusters and maintain open syllables, given that both consonant clusters and closed syllables (with the exception of syllable-final /n/) are not permitted in Japanese (e.g., Dupoux et al., 1999). When it comes to segment deletion, despite being less frequently observed than vowel insertion (see Shoji and Shoji, 2014), English inflectional suffixes or word-final consonants might be deleted to avoid closed syllables.

Alongside the three categories (SUB, INS, DEL) outlined above, an additional morpho-phonological category was introduced for boundary transgressions (BTR). This category contained Japanese adaptations resulting from a reanalysis of the morphological boundaries of an English phrase and thus subjecting it to Japanese morphology. For example, the English phrase [saɪts ɔn ju] ("sights on you") was perceived as the Japanese word [saiso:nʲɯ:] (saisōnyū, Eng. "reinsertion"), involving a reanalysis of the word-final consonant /s/ in English [saɪts] as the onset of the second syllable in the Japanese misperception.
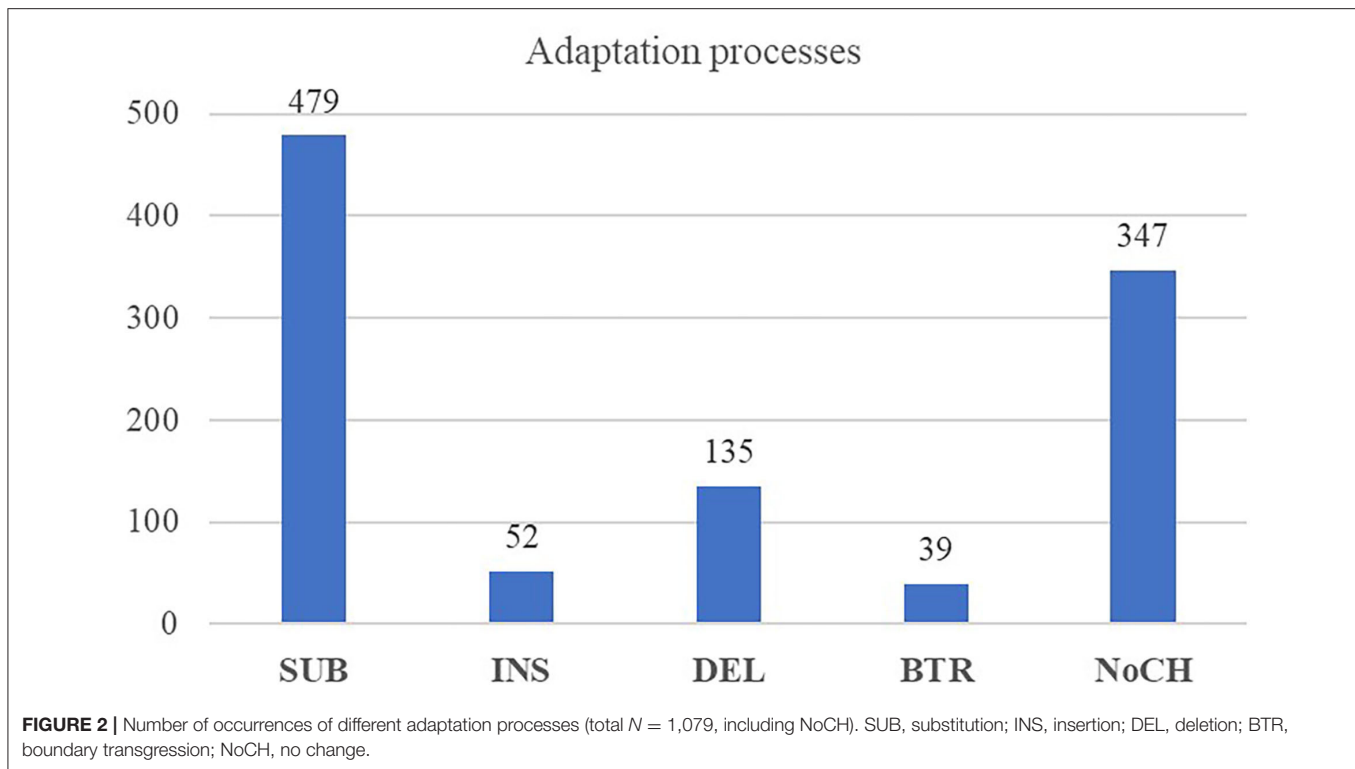
A fifth category labeled "no change" (NoCh) was included for sounds or sound sequences which did not reveal any phonological or morphological differences between the source lyrics and the misperceived lyrics. **Table 2** gives an example of a *soramimi* taken from our corpus, including the English source lyrics, the Japanese misperceived lyrics and the corresponding IPA transcriptions. The phonological processes identified in this example are color-coded.

ANALYSIS

As displayed in **Figure 2**, out of the total number of phonological adaptation processes identified in the corpus ($N = 1,079$), the largest number is constituted by the process of sound substitution ($N = 479$), i.e., by the replacement of English sounds by sounds that are part of Japanese phonology.

Oftentimes, SUBs affected vowels or consonants that are part of the English sound inventory, but do not exist in the Japanese inventory, while in some other cases sounds that do exist in Japanese were replaced by other Japanese sounds. For example, the English phrase [hoʊp kæn kip mi tugɛðə] ("Hope can keep me together") was misheard as [hokeŋkiN mitsɯketa] ("Hokenkin mitsuketa"—"I found the insurance money"). In this mishearing, the following substitutions can be observed:

Vowels (Eng. → Jpn.): oʊ →o / a →e / u →ɯ/ ɛ →e / ə→a

---

[5]We use the term *English-Japanese mondegreens* to refer to English song lyrics which have been adapted to the Japanese language by native Japanese listeners.
[6]As "felicitous" we labeled such *soramimi* to which the show's hosts gave high ratings and reacted positively and where therefore perception of the mishearing was arguably shared by them.

FIGURE 2 | Number of occurrences of different adaptation processes (total *N* = 1,079, including NoCH). SUB, substitution; INS, insertion; DEL, deletion; BTR, boundary transgression; NoCH, no change.

Consonants (Eng. → Jpn.): n →ŋ / p →N / t →ts / g →k / ð →t

In the case of the vowels, there are four that do not exist in Japanese, namely /oʊ/, /æ/, /ɛ/ and /ə/, all of which were replaced by vowels that are the closest approximations in the Japanese vowel inventory. The vowel /u/ itself is not part of the Japanese vowel inventory, but a similar sound, [ɯ], is, which is why it substitutes for /u/. As for the consonants, while /n/ is also phonemic in Japanese, when positioned in a syllable coda it is realized as a moraic nasal /N/ and is known to assimilate to the following sound (Vance, 2008) as in the example above. This is also what happens in the case of /p/, where the replacement with Japanese /N/ occurs because it precedes a bilabial nasal /m/ and thus becomes subject to nasalization. The plosive /t/, again, exists in Japanese, but is restricted with regard to the vowels it can combine with. While the syllables /ta/, /te/, and /to/ are licensed, /tu/ and /ti/ are not. For these two cases, /t/ is replaced by Japanese affricates, i.e., /ts/ combines with /u/, and /tɕ/ with /i/. As a consequence, English /tu/ is replaced in this mishearing with /tɕɯ/. Finally, the English fricative /ð/ is another unknown consonant in Japanese, and is thus substituted by the alveolar stop /t/, which is a common process in languages that do not contain interdental fricatives (Blevins, 2006).

Another example is the mishearing of [wəgʊɾ ɪz əkɪs] ("What good is a kiss") as [ɰakaɾe zakessɯᵝ] ("Wakaresakessu"—"It's a goodbye drink"). In this example, the following substitutions can be identified:

Vowels (Eng. → Jpn.): ə→ a (2x) / ʉ → a / ɪ → e (2x)

Consonants (Eng. → Jpn.): w → ɰ /g → k

As in the previous example, the English schwa sound /ə/ is replaced by /a/ in the Japanese mishearing, as is the English close central rounded vowel /ʉ/. While there is only a superficial phonetic similarity between /ʉ/ and /a/ in that they are both unrounded, it can be argued that this misperception is construed to render this sequence meaningful. This means that, in this particular case, it may be that top-down processing, the impetus to make and perceive a real word, overrides the lack of similarity between the original vowel and the perceived vowel. The vowel /ɪ/, in this example, is twice perceived as an /e/-vowel by the Japanese listener, even though there is a more similar vowel, namely /i/, available in the Japanese inventory. However, it has to be noted that the vowel /ɪ/ in the original English lyrics, especially for /ɪz/, is rather centralized and indeed could be considered to be somewhere between /ɪ/ and /ə/, which might explain the substitution. The two consonant changes are essentially substituting the non-phonemic sound /w/ with its closest approximation in Japanese, i.e., /ɰ/, and changing the voicing of the velar stop (/g/ to /k/), possibly because the background music makes it difficult for listeners to perceive whether the stop is voiced or voiceless, so place and manner of articulation in the previous example are retained while voicing changes in the Japanese misperception. In addition, stops in Japanese, according to Fujimoto, are "more aspirated word-initially than word-medially" (Fujimoto, 2015, p. 201) and thus the distinction between voiced and voiceless stops may be more difficult to draw in medial positions, as in this case the velar stop /k/ in "wakare".

As a last example for substitutions, the following mishearing will be discussed: [dontʃɑ nid səm lʌv tənaɪt˥] ("Don't you need some love tonight"), which was perceived as [to:tɕaN nʲi:tɕaN jotte nai] ("Tōchan, niichan yottenai"—"My dad and older brother are not drunk"). The following sound substitutions can be observed in this example:

Vowels (Eng. → Jpn.): ə→ a / ʌ → o / ə→ e
Consonants (Eng. → Jpn.): d → t / tʃ → tɕ / ds → tɕ / m → N / l → j

Once more, it can be observed that the English schwa sound, not being part of the Japanese vowel system, was perceived as /a/ and /e/, respectively, by the Japanese listener. The corpus data here shows that /a/ is the substitute in the large majority of cases, except for those where the schwa is preceded by /t/ or /θ/. The vowel /ʌ/ in "love" is replaced by the closest vowel in the Japanese vowel inventory, which is the vowel /o/. For the consonants, word initial /d/ is substituted by its voiceless counterpart /t/, possibly so perceived because of the increased aspiration in /d/ in the sung performance. There are two substitutions with the Japanese affricate /tɕ/: In the case of English /tʃ/, the Japanese affricate /tɕ/ is the most similar sound and thus substitutes for the English affricate. On the other hand, in the case of English /ds/, arguably through voicing assimilation, the /d/ might be more similar to /t/. In Japanese, this would create the illicit sound sequence /tsə/ or—after substitution of /ə/ by /a/—/tsa/; hence, it is substituted with the closest Japanese sound sequence /tɕa/. Finally, an interesting aspect is the omission of the /n/ sound in /dontʃɑ/, which seems to reappear later, at the end of the sequence in /to:tɕaN/, which we would argue indicates a perseverative effect of the nasal.

While a considerable amount of sound substitutions was identified in our corpus, insertions were comparatively rare (*N* = 52). This seems reasonable as the addition of non-existent vowels would alter the rhythm of the song lyrics and would therefore make it less likely to be perceptually shared by others. Consequently, most of the insertions identified in our corpus relate to those necessary in Japanese phonotactics, i.e., inserting vowels to break up consonant clusters. In the following, we will discuss three examples of such insertions.

Example 1:
Eng. [nɑːθɪn kan howld ʌs an:ɑθɪn kan kip ʌsdawn] ("Nothing can hold us and nothing can keep us down") → Jpn. [nadze ka hon daɕita, nadze ka ɕitto ɕita] ("Naze ka hon dashita, naza ka hitto shita"— "For some reason, I published a book and for some reason, it became a hit")
Example 2:
Eng. [kɪdz kʊdn hə dʒæk ðeɪ tɹaɪ tɹaɪ tɹaɪ] ("Kids couldn't hurt Jack. They tried and tried and tried") → Jpn. [k i tsɯᵝkɯᵝ ΦɯᵝN dʑa itai itai itai] ("Kitsuku hunja itai itai itai"— "It really hurts when you step on it")
Example 3:
Eng. [(ð)ə naɪ dej dɹowv owl dɪksi dawn] ("The night they drove old Dixie down") → Jpn. [inai dze doɾobo: dekʲiɕi da] ("Inaize dorobo. Dekishi da."—"The robber's gone. He drowned")

In Example 1, we can see that the vowel /i/ was inserted twice. While in the second case, it is inserted to break up what would otherwise be an illicit consonant cluster (/ds/), in the first instance, the situation is more complex as the original lyrics already include a licit sound combination, namely /sa/. Since the case for insertion to preserve Japanese phonotactics cannot be made here, it may be argued that the mishearing up to this point ("hon das") primes the listener to mentally complete what is an established idiom in Japanese ("hon dashita"— "published a book"), which would be another example of top-down processing.

In Example 2, the epenthetic vowel [ɯᵝ] is used to break up the English consonant cluster /dzk/. Japanese /N/, as we have argued before, may have been perceived and thereby inserted, due to the perseverative effect of the—otherwise deleted—/n/ in the word /kʊdn/, i.e., the sound appears later in the sequence in the mishearing by way of metathesis.

Example 3 is another case where consonant clusters are altered to make the sequence adhere to Japanese phonotactics. The original lyrics include two clusters, /dr/ in "drown" and /ks/ in "Dixie", which are modified in the mishearing by inserting the epenthetic vowel /o/ after /d/, and /i/ after /k/.

Out of the 135 deletions identified in the corpus, *N* = 99 were consonant deletions, showing that consonants were far more likely to be deleted than vowel targets. In *N* = 15 out of these deletions, word-final, mainly unreleased English plosives were omitted in the Japanese misperceptions, as shown in the following examples:

Example 4:
Eng. [bæk ɪnə həʊl weə ðeɪ gɑt˥] ("Back in the hole where they got") → Jpn. [gakʲi no ho: ɯa dekai] ("gaki no hō wa dekai"—"the little one's is bigger")
Example 5:
Eng. [dontʃɑnidsəmlʌvtənaɪt˥] ("Don't you need some love tonight") → Jpn. [to:tɕan nʲi:tɕaN jotte nai] ("Tochan, niichan yottenai"—"My dad and older brother are not drunk")
Example 6:
Eng. [hoʊp˥ kan kip mi tugɛðə] ("Hope can keep me together") → Jpn. [hokeŋkʲiN mʲitsɯᵝketa] ("Hokenkin mitsuketa"— "I found the insurance money")

In these examples, the post-vocalic alveolar plosive [t˥], lacking an audible release burst in the English syllables [gɑt˥] and [naɪt˥], and the unreleased word-final English bilabial plosive [p˥] in [hoʊp˥] were deleted in the corresponding Japanese misperceptions. A possible phonetic explanation for this is the observation that syllable-final plosive consonants which are produced without an audible release are usually less intelligible than released plosives (e.g., Lisker, 1999; Nozawa and Cheon, 2014). This is particularly the case if the plosive target is preceded by a diphthong, as in Examples 5 and 6 above, potentially resulting from within-vowel formant movements which may impede listeners' ability to disentangle consonant and vowel components (see Lisker, 1999). Alongside phonetic considerations, deleting unreleased final plosives is likely to be a consequence of subjecting English input to Japanese phonotactic

constraints. In Japanese, the plosives /p b t d k g/ exclusively occur in syllable-initial position given that closed syllables are disallowed (e.g., Ito and Mester, 2015). Hence, it can be argued that deleting syllable-final English consonants—whether they are released or not—aims to maintain open syllables, as shown in the above examples. This is, of course, not restricted to final plosive consonants, but may also affect other English consonants occurring in syllable-final position and thus constituting closed syllables. For instance, the final fricative /v/ in the English lyrics [aɪsɛdawkʊːt hi ɛvəliːv] ("I said 'How could he ever leave'") was deleted in the Japanese misperception [asedakɯᵝ çiabɯᵝrʲi] ("Ase daku hi aburi"—"Sweating and burning") to maintain an open syllable.

Similar observations were made with regard to English consonant clusters where the second consonant of the cluster was omitted in the Japanese misperception. In Example 2 above, the word [tɹaɪ] in the English lyrics [kɪdz kʊdn həd͡ʒæk ðeɪ tɹaɪ tɹaɪ tɹaɪ] ("The kids couldn't hurt Jack, they tried, tried, tried") was reinterpreted as [itai] ("it hurts") in the Japanese misperception [k i tsɯᵝkɯᵝ ΦɯᵝN d͡za itai itai itai], involving a deletion of the English approximant [ɹ] and thus avoiding the clustering of two successive consonants in Japanese.

In terms of segment deletions, it was further observed that diphthongs occurring in the English source lyrics were frequently perceived as monophthongs in Japanese:

Example 7:
Eng. [ʔɪdoʊn mɛɪkɪʔizei] ("Ya don't make it easy") → Jpn. [ittɕoːme de iː dze] ("Icchōme de ii ze"—"You can drop me off after just one block")

Example 8:
Eng. [meɪ teja gawd ɔmateɪ goʊna kʊtʃuda] ("[Let] me tell you God Almighty gonna cut you down") → Jpn. [mʲi te jo, koɾe marʲiko no kɯᵝtsɯᵝ da] ("Mite yo kore Mariko no kutsu da"—"Look, it's Mariko's shoes")

In these examples, the English diphthongs /oʊ/, /ɛɪ/, and /eɪ/ were reinterpreted as the Japanese monophthongs /o/, /e/, and /i/, respectively, in each case involving the deletion of one of the two adjacent vowel targets in the dipthongs. As discussed in Section Integration Processes for Foreign Language Material in Japanese, the Japanese vowel inventory does not generally contain diphthongs (with a few exceptions only, see e.g., Kubozono, 2005) and, hence, native Japanese listeners might encounter difficulties perceiving English diphthongs and replace them with monophthongs instead.

As shown in **Figure 2**, the lowest number of processes identified in the speech corpus is constituted by boundary transgressions (*N* = 39), that is, reanalyzing morphological boundaries of an English phrase to adapt it to Japanese morphology. As illustrated in the following examples, boundary transgressions were identified in cases where an English consonant in coda-position was perceived as the syllable-onset of the following word in Japanese:

Example 9:
Eng. [jɤːdownowmaneː] ("You don't know my name") → Jpn. [joɯᵝ doɯᵝ nomanee] ("Yo! Do! Nomanee!"—"Hey, how about a drink?")

Example 10:
Eng. [saɪts ɔn ju] ("sights on you") → Jpn. [sai soːnʲɯᵝː] ("saisōnyū"—"reinsertion")

Example 11:
Eng. [ɛvɹidaɹkɛskaj] ("Every darkest sky") → Jpn. [ebʲi dake ssɯᵝ ka] ("Ebi dake ssuka"—"Only shrimp?")

In Example 9, the coda consonant in English [down] was perceived as the onset consonant in the Japanese word [nomanee]. Similarly, in Examples 10 and 11, syllable-final English /s/ in [saɪts] and [daɹkɛs] was reanalyzed as the onset consonant in the Japanese targets [sai soːnʲɯᵝː] and [ssɯᵝka], respectively. Example 10 additionally shows a deletion of the English alveolar plosive /t/ preceding the coda consonant /s/, which allows for the open syllable [sai] in the Japanese misperception.

DISCUSSION AND CONCLUSION

As our analysis has shown, the processes that underlie the majority of perceptual changes between the original English song lyrics and their Japanese misperceptions largely follow the patterns also observable in the perception of English in general and of English loanwords in particular (e.g., Loveday, 1996; Stanlaw, 2004). This confirms our initial hypothesis that the perception of foreign language input is likely to be filtered through and affected by L1 phonotactics and hence subject to the same processes as loanword integration *via* the auditory channel.

In essence, the reinterpretation of English auditory input occurs through the phonological filter of Japanese, which is reflected in the frequent occurrence of various phonological and morphophonological processes, namely sound substitutions, insertions, segment deletions, and boundary transgressions. All these processes are aimed at making the perceived input adhere to Japanese phonotactics by maintaining open syllables, breaking up consonant clusters by means of epenthetic vowels as well as by reanalyzing word-final consonant syllable codas as syllable onsets of the following phrase (see e.g., Dupoux et al., 1999; Kubozono, 2002).

The bulk of evidence for perceptual assimilation is reflected in sound substitutions, which by far make up the largest number of phonological processes identified in our corpus. As discussed earlier in this paper, perceptual assimilation, as defined by Peperkamp et al. (2008, p. 131), describes "a process that applies during speech perception and that maps non-native sound structures onto the phonetically closest native ones". The examples in our corpus have shown that English sounds that are non-existent in Japanese were systematically, though not exclusively, replaced by the closest Japanese sounds. This

occurred not only for individual segments that are not part of the Japanese sound inventory, such as the dental fricatives /θ/ and /ð/ or the lateral approximants /r/ and /l/, but also for sound combinations that are not licit in Japanese, such as consonant clusters or certain CV-combinations, for example English /tu/ or /si/ (see e.g., Dupoux et al., 1999). In some particular cases, we identified English sounds being replaced by non-similar Japanese sounds, such as [ɯ] being substituted by [a]. For these cases, we argue that they might represent examples of top-down processing, which, as outlined in Section Cues to Processing Connected Speech, refers to listeners' ability to create meaningful linguistic units out of unclear auditory input (e.g., Warren, 1970; Ganong, 1980). When it comes to the perception of English song lyrics, this means that the listeners' instinct to perceive a meaningful word or phrase through their own language filter is so strong that this overrides the lack of phonological similarity between the original and the perceived segment. This could be seen in cases where the expectation of a certain word or phrase was already primed, thus arguably leading the listener to falsely perceive a sound despite a complete absence of similarity between the original English and the misperceived Japanese sound. In other cases, we found that voiced stops occurring in initial position in the English original would be substituted by voiceless stops occurring in medial position in the Japanese misperception, which we attributed to a lack of aspiration in stops in medial position in Japanese (see e.g., Fujimoto, 2015), making the voicing distinction more difficult to draw. Additionally, it might be argued that the misperception of plosive voicing results from adverse listening conditions, such as a high level of background noise, which has previously been shown to affect the accurate perception of English plosives (Brent et al., 2013).

As for insertions, in line with what we expected, these were less frequently observed. The reason for this, we argue, is that including additional vowels would not only extend the phrase in terms of syllable number, but would also thereby break up the original rhythm of the song, making the reinterpretation less likely to be perceived, and/or shared by others. Where insertions did occur, this was mostly to break up consonant clusters in the English original, i.e., to make the mishearing adhere to Japanese phonotactics. In other, rarer cases, additional syllables were inserted to create a meaningful utterance or complete an idiom that was perceived, priming the listener to it, but where the final syllable of that idiom was not part of the original lyrics. Finally, an interesting phenomenon we identified was the occasional insertion of /N/ which occurred at a different position in the original. The reason, we argue, why this nasal was perceived at a later position in the mishearing by way of metathesis is due to both a perseverative effect of the nasal and due to top-down processing, in which the expectancy to hear a meaningful word led to the reassignment of the sound to a different position.

Deletions served a similar purpose to insertions, namely to maintain open syllables, by deleting sounds such as coda consonants. In many cases, these consonants were unreleased English voiceless plosives which are less easily perceivable than plosives produced with an audible release (e.g., Lisker, 1999). In other cases, deletion was applied to consonant clusters as an alternative to epenthetic vowel insertion which, as discussed

above, would have interfered with the rhythm of the original lyrics. Additionally, the input being aural, final consonant deletion was an anticipated outcome, as unvoiced consonant endings may lead to a "lack of perception" (Shoji and Shoji, 2014, p. 9).

The last process, boundary transgression, was the most systematic in the sense that it exclusively occurred when an English word ended with a consonant coda, which was then—instead of omitting it—reanalyzed as the onset of the subsequent word or phrase in the Japanese mishearing. In some cases, this was facilitated by the following phrase having a vowel onset while in other cases, boundary transgressions combined with insertions of epenthetic vowels to ensure licit Japanese sound combinations.

As the above discussion shows, the last three processes—insertion, deletion, and boundary transgression—serve a common purpose, namely to maintain the legality of sound combinations in the misperception by either adding a sound, eliding a sound, or shifting the morphological boundaries of the original lexeme to ensure phonotactic compatibility. Simultaneously, substitution works to adapt the input to the Japanese sound inventory. As such, these four processes condition, affect and trigger each other in complex ways and therefore effectively reflect the impact of top-down processing and perceiving foreign language input through the listeners' L1 filter. We argue that the choice of the process to ensure adherence to Japanese phonotactics is clearly a factor of priming and expectancy bias, i.e., which word or phrase is anticipated, and that therefore, while seemingly random, whether a sound gets inserted, deleted or morphologically reanalyzed is a direct consequence of top-down processing.

While the findings obtained in the present analysis provide insights into the processes governing the misperception and reinterpretation of foreign language input, it needs to be taken into consideration that our speech corpus only included "felicitous" *soramimi*, that is, those that were immediately reproducible to the show hosts upon playing the clips. There are many other mishearings presented in the show which the hosts struggled with or found impossible to reproduce, hence it is not possible to generalize the present findings to include all *soramimi*, as some are clearly triggered by factors related to the listener and the listening situation. It is also possible that there was some editorial influence or that some of the examples were fabricated for the show; however, even if this did occur, the examples need to be similar enough to the English lyrics to be plausible to the audience. For the present analysis, we selected the best-practice examples in order to understand not merely the processes underlying crosslinguistic mondegreens, but also to get an idea of the conditions that need to be met for a mondegreen to be perceptually shared by other members of the speech community as well. While we cannot assume that our findings are generalizable to naturally-occurring data, the current analysis clearly shows patterns in terms of what processes are used and which ones are used more often.

Another potential limitation of a study of this kind, as addressed by Otake (2007), is that there is no information available concerning the listeners' linguistic background, that is,

whether they are, for instance, proficient speakers of English as a second language (L2) or (quasi-)monolingual Japanese speakers. This is of importance because native Japanese listeners who are highly proficient speakers of L2 English might be more likely to perceive the English lyrics accurately compared to less proficient speakers of L2 English (see e.g., Kaneko, 2006, for a discussion of the role of L2 proficiency on vowel perception).

Finally, the findings of this study are also limited in the sense that it is not possible to determine the specific environment in which the mishearing occurred and whether there were potential factors involved that facilitated the triggering of the misperception, such as music played in a club, on the radio, or in a car with multiple background noises (see Lecumberri et al., 2010, for a review of studies examining the effects of adverse listening conditions on the perception of non-native speech). Nevertheless, the current findings could be used as a starting point for further work involving controlled experiments with listeners of varying linguistic backgrounds as well as different listening contexts, in order to tease apart these influences.

As mentioned in Section Cues to Processing Connected Speech, priming through orthography and visuals might also play an important role in the misperceptions presented in *Soramimi Hour*. A potential follow-up study, therefore, could be to play English song lyrics used on *Soramimi Hour* to native Japanese listeners in various iterations: Only the music, with subtitles, with images but without subtitles, then with images and subtitles, and to compare whether and to what extent each of these factors contributes to the success of the mishearing and hence, to what extent such primes "manipulate" the perception of the participants. This would help to determine the specific contributions of the different factors (lyrics, subtitles, images) to the inducing of a particular reinterpretation.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

JS: study design, transcription and analysis of data, and manuscript writing. LK: analysis of data, manuscript writing, and tables and figures. NK: transcription and analysis of data and contributed to, reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica.* 66, 46–63. doi: 10.1159/000208930

Balmer, F. (2007). *Phonetic, Phonological and Prosodic Triggers for Mondegreens.* Unpublished manuscript. Available online at: http://www.ambiguitaet.uni-tuebingen.de/minigrad/balmer/Balmer_(2007)_Phonetic,_Phonological_and_Prosodic_Triggers_for_MondegreensFINAL.pdf (accessed October 20, 2021).

Beck, C., Kardatzki, B., and Ethofer, T. (2014). Mondegreens and soramimi as a method to induce misperceptions of speech content: influence of familiarity, wittiness, and language competence. *PLoS ONE* 9, e84667. doi: 10.1371/journal.pone.0084667

Blevins, J. (2006). New perspectives on english sound patterns: 'natural' and 'unnatural' in evolutionary phonology. *J. Engl. Linguist.* 34, 6–25. doi: 10.1177/0075424206287585

Bolinger, D. (1965). "Pitch accent and sentence rhythm," in *Forms of English: Accent, Morpheme, Order,* eds. T. Kanekiyo and I. Abe (Harvard: Harvard University Press), 139–80.

Brent, J. A., Patra, H., and Vaz, P. C. (2013). Effects of noise and speaker's language background on plosive perception. *J. Acoust. Soc. Am.* 134, 4229. doi: 10.1121/1.4831538

Chéreau, C., Gaskell, M. G., and Dumay, N. (2007). Reading spoken words: orthographic effects in auditory priming. *Cognition* 102, 341–360. doi: 10.1016/j.cognition.2006.01.001

Dame, M. A., Patra, H., and Vaz, P. C. (2013). Effects of speakers' language background on speech perception in adults. *J. Acoust. Soc. Am.* 134, 4063. doi: 10.1121/1.4830833

Damian, M. F., and Bowers, J. S. (2009). Assessing the role of orthography in speech perception and production: evidence from picture–word interference tasks. *Eur. J. Cogn. Psychol.* 21, 581–598. doi: 10.1080/09541440801896007

Dower, J. W. (1999). *Embracing Defeat.* New York, NY: W. W. Norton and Company.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J. (1999). Epenthetic vowels in Japanese: a perceptual illusion. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 1568. doi: 10.1037/0096-1523.25.6.1568

Fuchs, R. (2016). *Speech Rhythm in Varieties of English: Evidence from Educated Indian English and British English.* Singapore: Springer.

Fujimoto, M. (2015). "4 Vowel Devoicing," in *Handbook of Japanese Phonetics and Phonology,* ed. H. Kubozono (Berlin: De Gruyter Mouton), 167–214. doi: 10.1515/9781614511984.167

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *J. Exp. Psychol. Hum. Percept. Perform.* 6, 110–125. doi: 10.1037/0096-1523.6.1.110

Gick, B., and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature.* 462, 502–504. doi: 10.1038/nature08572

Gordon, R. L., Schön, D., Magne, C., Astésano, C., and Besson, M. (2010). Words and melody are intertwined in perception of sung words: EEG and behavioral evidence. *PLoS ONE* 5, e9889. doi: 10.1371/journal.pone.0009889

Guion, S. G., Flege, J. E., Akahane-Yamada, R., and Pruitt, J. C. (2000). An investigation of current models of second language speech perception: the case of Japanese adults' perception of English consonants. *J. Acoust. Soc. Am.* 107, 2711–2724. doi: 10.1121/1.428657

Hagiwara, R. (1997). Dialect variation and formant frequency: the American English vowels revisited. *J. Acoust. Soc. Am.* 102, 655–658. doi: 10.1121/1.419712

Hayashi, M., and Kakehi, K. (1990). "An experimental study on basic perceptual units of speech based on reaction time," in *Paper Presented at the Spring Meeting of the Acoustic Society of Japan.*

Hoffer, B. L. (2002). "The impact of English on the Japanese language," in *Exploring Japaneseness: On Japanese Enactments of Culture and Consciousness,* ed. R. T. Donahue (Westport: Greenwood Publishing), 263–73.

Hogg, R. M. (1992). "Phonology and morphology," in *The Cambridge History of the English Language,* ed. R. M. Hogg (Cambridge: Cambridge University Press), 67–167.

Inoue, H. (1996). *Nihongo Nikki, Vol. II (Diary of Japanese, Vol. II).* Tokyo: Bungei Shunjuu.

Ito, J., and Mester, A. (2015). "7 Sino-Japanese phonology," in *Handbook of Japanese Phonetics and Phonology,* ed. H. Kubozono (Berlin: De Gruyter Mouton), 289–312.

Johnson, K., Strand, E. A., and D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *JPhon* 27, 359–384. doi: 10.1006/jpho.1999.0100

Jungers, M. K., Palmer, C., and Speer, S. R. (2002). Time after time: the coordinating influence of tempo in music and speech. *Cog. Proc.* 1, 21–35.

Kaneko, E. (2006). Vowel selection in Japanese loanwords from English. *LSO Working Papers in Linguistics* 6, 49–62.

Kentner, G. (2015). "Rhythmic segmentation in auditory illusions-Evidence from cross-linguistic mondegreens," in *Proceedings of 18th ICPhS*. Glasgow.

Kubozono, H. (2002). "Mora and Syllable," in *The Handbook of Japanese Linguistics*, ed. N. Tsujimura (Oxford: Blackwell), 31–61.

Kubozono, H. (2005). [ai]-[au] asymmetry in English and Japanese. *J. Eng. Linguist.* 22, 1–22. doi: 10.9793/elsj1984.22.1

Kubozono, H. (2015a). "Introduction to Japanese Phonetics and Phonology," in *Handbook of Japanese Phonetics and Phonology*, ed. H. Kubozono (Berlin: Walter de Gruyter), 54–93.

Kubozono, H. (2015b). *Handbook of Japanese Phonetics and Phonology*. Berlin: Walter de Gruyter.

Labrune, L. (2012). *The Phonology of Japanese*. Oxford: Oxford University Press.

Ladefoged, P. (2005). *Vowels and Consonants: An Introduction to the Sounds of Languages*. Malden, MA: Blackwell Pub.

Ladefoged, P., and Broadbent, D. E. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98–104. doi: 10.1121/1.1908694

Ladefoged, P., and Johnson, K. (2014). *A Course in Phonetics*. Wadsworth: Cengage Learning.

LaFeber, W. (1997). *The Clash - U.S.-Japanese Relations throughout History*. New York, NY: Norton.

Lecumberri, M. L. G., Cooke, M., and Cutler, A. (2010). Non-native speech perception in adverse conditions: a review. *Speech Commun.* 52, 864–886. doi: 10.1016/j.specom.2010.08.014

Lisker, L. (1999). Perceiving final voiceless stops without release: effects of preceding monophthongs versus nonmonophthongs. *Phonetica* 56, 44–55. doi: 10.1159/000028440

Loveday, L. J. (1996). *Language Contact in Japan - a Socio-Linguistic History (Oxford Studies in Language Contact)*. Oxford: Oxford University Press.

Mann, V. A., and Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Percept. Psychophys.* 28, 213–228. doi: 10.3758/BF03204377

Matsuda, Y. (1985). Cross-over languages: Japanese and English (I). *Kwansei Gakuin Univ. Ann. Stud.* 34, 41–69.

Matsuda, Y. (1986). Cross-over languages: Japanese and English (II). *Kwansei Gakuin Univ. Ann. Stud.* 35, 47–80.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Morrow, P. R. (1987). The users and uses of English in Japan. *World Englishes* 6, 49–62. doi: 10.1111/j.1467-971X.1987.tb00176.x

Nian, O. S., and Jubilado, R. C. (2011). The linguistic integration of English borrowings in modern Japanese. *Polyglossia* 21, 99–107.

Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., and Trent-Brown, S. A. (2008). Acoustic and perceptual similarity of Japanese and American English vowels. *J. Acoust. Soc. Am.* 124, 576–88. doi: 10.1121/1.2931949

Nozawa, T., and Cheon, S. Y. (2014). The identification of stops in a coda position by native speakers of American English, Korean, and Japanese. *J. Phon. Soc. Japan* 18, 13–27. doi: 10.24467/onseikenkyu.18.1_13

Ohata, K. (2004). Phonological differences between Japanese and English: several potentially problematic. *Lang. Learn.* 22, 29–41.

Ōishi, I. (1992). Japanese attitudes towards the english language during the pacific war. *Seikei Hogaku* 34, 1–19.

Olson, I. A. (1997). The dreadful death of the bonny earl of murray: clues from the carpenter song collection. *Folk Music J.* 7, 281–310.

Otake, T. (2007). "Interlingual near homophonic words and phrases in L2 listening: Evidence from misheard song lyrics," in *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, Saarbrücken: University of Saarland, 777–780.

Otake, T., Yoneyama, K., Cutler, A., and Van Der Lugt, A. (1996). The representation of Japanese moraic nasals. *J. Acoust. Soc. Am.* 100, 3831–3842. doi: 10.1121/1.417239

Peperkamp, S. (2005). "A psycholinguistic theory of loanword adaptations," in *Proceedings of the 30th Annual Meeting of the Berkley Linguistics Society*, Berkeley, CA.

Peperkamp, S., Vendelin, I., and Nakamura, K. (2008). On the perceptual origin of loanword adaptations: experimental evidence from Japanese. *Phonology* 25, 129–164. doi: 10.1017/S0952675708001425

Pike, K. L. (1945). *The Intonation of American English*. Westport, CT: Greenwood Press.

Pike, K. L. (1947). *Phonemics - A Technique for Reducing Languages to Writing*. Ann Arbor, MI: University of Michigan.

Pintér, G. (2015). "The emergence of new consonant contrasts," in *Handbook of Japanese Phonetics and Phonology*, ed. H. Kubozono (Berlin: Walter de Gruyter), 172–214.

Scherling, J. (2012). *Japanizing English: Anglicisms and Their Impact on Japanese*. Tübingen: Narr Verlag.

Scherling, J. (2013). Holistic loanword integration and loanword acceptance. A comparative study of anglicisms in German and Japanese. *AAA-Arbeiten aus Anglistik Amerikanistik* 38, 37–52.

Shoji, S., and Shoji, K. (2014). "Vowel epenthesis and consonant deletion in Japanese loanwords from English," in *Proceedings of the 2013 Annual Meetings on Phonology*. Amherst, MA: University of Massachusetts, 1.

Stanlaw, J. (2002). Hire or fire? Taking AD-vantage of innovations in the Japanese syllabary system. *Lang. Sci.* 24, 537–574. doi: 10.1016/S0388-0001(01)00004-3

Stanlaw, J. (2004). *Japanese English - Language and Culture Contact*. Hong Kong: Hong Kong University Press.

Starr, R. L., and Shih, S. S. (2017). The syllable as a prosodic unit in Japanese lexical strata: evidence from text-setting. *Glossa* 2, 1–34. doi: 10.5334/gjgl.355

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *J. Lang. Soc. Psychol.* 18, 86–99. doi: 10.1177/0261927X99018001006

Tang, K. (2015). *Naturalistic speech misperception* (dissertation). University College London, London, United Kingdom.

Tsuchida, A. (1995). English loans in Japanese: constraints in loanword phonology. *Working Papers Cornell Phonet. Lab.* 10, 145–164.

Turk, A., and Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Lab. Phonol.* 4, 93–118. doi: 10.1515/lp-2013-0005

Vance, T. J. (1987). *An Introduction to Japanese Phonology*. Albany, NY: State University of New York Press.

Vance, T. J. (2008). *The Sounds of Japanese*. Cambridge: Cambridge University Press.

Venedlin, I., and Peperkamp, S. (2006). The influence of orthography on loanword adaptations. *Lingua* 116, 996–1007. doi: 10.1016/j.lingua.2005.07.005

Warner, N., and Arai, T. (2001). Japanese mora-timing: a review. *Phonetica*. 58, 1–25. doi: 10.1159/000028486

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science* 167, 392–393. doi: 10.1126/science.167.3917.392

Wassink, A. B. (2001). Theme and variation in Jamaican vowels. *Lang. Var. Change* 13, 135–159. doi: 10.1017/S0954394501132023

Wright, S. (1954). *The Death of Lady Mondegreen*. Harper's Magazine, 48–51.

Zec, D. (2007). "The syllable," in *The Cambridge Handbook of Phonology*, ed. P. de Lacy (Cambridge: Cambridge University Press), 161–94.

frontiers | Frontiers in Communication

The Introduction of English-Induced Neologisms in Spanish Tweets: A Case-Study on *covidiota*

Franziska Kailich*

Department of Romance Linguistics, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

This paper intends to trace the introduction of an English-induced, COVID-related neologism, *covidiota*, into the Spanish language. The study is based on a corpus of tweets, starting in March 2020. It examines several specific features which mark the word as a new, unfamiliar item, such as different ways of graphical highlighting, for example. On the other hand, the paper aims to detect possible indicators of an integration of *covidiota* into the Spanish language use in the tweet corpus compiled for this case study.

Keywords: language contact, lexical innovation, semantic change, social media, verbal humor

INTRODUCTION

The 2020 pandemic caused by the outbreak of COVID-19 has had an impact on a great many aspects of life. Amongst other things, it has brought about the creation of a number of neologisms (cf., e.g., McPherson et al., 2020). The Spanish word *covidiota* is a particularly noteworthy example of such a COVID-related innovation: it is an anglicism, and a blend. Blending is typically considered a marginal phenomenon amongst word-formation processes, in Spanish as well as in other languages (cf., e.g., Pharies, 1987, p. 271). Among frequently named characteristics of blends, there is a humorous, jocular note these words convey (e.g., Lang, 1990, p. 199; Varela Ortega, 2005, p. 95).

This paper intends to trace the development of Span. *covidiota* as an example of a neologism induced by language contact, which creates its own dynamism within the receiving language. The situation of language contact in this specific case arises in online communication on social media platforms. Language use on the Internet is often said to increase the linguistic influence of English on languages worldwide (e.g., Balteiro, 2012, p. 9). As Twitter represents a particularly useful source for the linguistic analysis of online language use (Friginal et al., 2018, p. 342–345), the case study on Span. *covidiota* is based on the analysis of tweets, starting in March 2020.

During the first week of its existence on Twitter, the word *covidiota*, when used in Spanish tweets, frequently appears together with an English definition of the term *covidiot*, sometimes commented on in Spanish or accompanied by a translation or additional definition in Spanish. But while the early instances of *covidiota* in Spanish tweets are frequently limited to citing and defining the new word, tweets from later periods show the term in use.

A certain collective negotiation of the semantic content remains a prominent feature in the tweets, which frequently revolve around the question of when or why a person or an action qualifies as a *covidiota*. The basic meaning could be given as 'someone who behaves foolishly with regard to the COVID-19 pandemic' (or, since Span. *covidiota* can also be used as an adjective, 'behaving in a foolish way regarding the COVID-19 pandemic'). Beyond that, there seems to be a continuum of uses that stretches from 'jocular term used playfully to refer to oneself or someone else' to 'swearword, used to insult and ridicule.' A very noticeable finding is that the term is used by both supporters and opponents of antivirus measures, designating members of the other group, respectively.

The main focus of the paper is on the introduction of the term into the Spanish language and several sub-processes involved in it. In order to describe these processes, several research parameters are established. The phenomena under investigation can be described and grouped together as markers of newness or otherness, on the one hand, and indicators of integration into the receiving language, on the other. This does not mean that there are two kinds of processes at work which operate in opposite directions. The study rather aims at describing two sides of one process: a decreasing presence of markers of newness is assumed to go hand in hand with a growing presence of indicators of integration.

THEORETICAL AND TERMINOLOGICAL FOUNDATION OF THE CASE STUDY

Blending in English and Spanish

The English term 'blending' refers to a specific word formation process by which two (or, in very few cases, more than two) words are merged together to form a new word, a blend. Blends can thus be defined as "two-constituent compounds in which at least one constituent has lost some phonological material" (Arndt-Lappe and Plag, 2013, p. 5). Although in cases such as *covidiota* both constituents *covid* and *idiota* are contained in the blend, the overlapping phonological material (the syllable *id* in this case) is not repeated in both constituents (*covididiota*), but shared between them (*covidiota*). This differentiates blends from compounds, which also constitute a combination of two (or more) lexical units. However, the constituents usually maintain their form and are largely unchanged in the process of compounding. The amount of material that gets lost in blending can differ and depends on whether the constituent words share a sound or sequence of sounds. If they do, the merging usually takes place within that sequence (Ulašin, 2016, p. 164). As a consequence, when morphologically analyzing a blend, it becomes obvious that the constituents do not usually coincide with morpheme boundaries. In blending, matters of prosody, syllabic structure and accent, are of a higher priority than morphological accuracy.

Blending is frequently considered a marginal word formation type not only because of its disregard for morpheme boundaries, but because it is not a very frequent phenomenon (Pharies, 1987, p. 271). To be more precise, a relatively small number of blends only ever become widely known within a speech community, meaning that most words that are formed by means of this particular process never become lexicalized. In spite of its alleged marginality, blending is a relevant mechanism of lexical innovation (Ulašin, 2016, p. 181), mainly because of another very distinctive quality: Blending is a very resourceful mechanism for forming particularly creative neologisms. Apart from meeting the need for linguistic economy (since it enables speakers to express a complex meaning in a condensed form), blending is generally perceived as conveying a certain nuance of playfulness, and is therefore very often and very characteristically linked to a humorous intention (Pharies, 1987, p. 281; Lang, 1990, p. 199; Varela Ortega, 2005, p. 95; Ulašin, 2016, p. 179).

Several authors, albeit without referring to any quantitative, comparative studies of the matter, point out that in English, blending has been much more common than in other languages (e.g., Lang, 1990, p. 199–200; Ulašin, 2016, p. 181) and that an increasing number of blends in several other languages can be traced back to the influence of English on these languages (Roig-Marín, 2017, p. 52). Balteiro (2018, p. 2) points out that "due to English influence blending is evolving from a minor device in Spanish into an important one, not only to cover lexical needs but also for creative and stylistic reasons (to produce humor, irony, punning, etc.)". When it comes to describing the introduction of *covidiota* into Spanish, this aspect is relevant because it helps to explain the conspicuousness of the neologism with regard to the Spanish language system: *covidiota* is a blend, i.e., created by means of a word formation process which stands out in its own right, for the reasons described above. Moreover, *covidiota* is a blend which was first coined in English and only then introduced into Spanish.

Language Contact in Social Media Communication

Language on the internet, and thus also language in social media-communication, is in many regards subject to a strong influence from English (e.g., Balteiro, 2012, p. 9; Balteiro, 2018, p. 2), even though the internet has definitely stopped being an all-English medium and has become a multilingual one (Crystal, 2011, p. 78). Following Kachru's categorization, the Spanish speaking Twitter users whose language usage is studied in the present case study can be said to belong to the "Expanding circle," which is defined by people who learn and use English as a foreign language (Kachru, 1994, p. 136). It is, however, also perfectly possible for at least a part of them to belong to the "Outer Circle," or even to the "Inner Circle" (Kachru, 1994, p. 136), depending on whether they use English as a second or as their first language. The tweet corpus compiled for the present study comprises tweets from all over the Spanish speaking world, and it is likely that it includes several texts composed by bilingual speakers of English and Spanish, for example in the USA.

Apart from bilingualism as an influencing factor originating from the speakers (Twitter users), the language used in digital communication can generally be said to possess certain characteristics in its own right, which, again, include a strong influence of English. Balteiro depicts this kind of language as basically a "new and rapidly-changing and developing language as a result of a rapid evolution or adaptation to Internet requirements of almost all world languages, for whom English is a trendsetter" (2012, p. 9). She adapts David Crystal's term *Netspeak* (Crystal, 2004, p. 78; Crystal, 2006, p. 19–25) and thus refers to the variety of Spanish used in digital communication as *Spanish Netspeak* (Balteiro, 2012, p. 10).

It can be assumed that both aspects, i.e., the bi- or multilingualism of Twitter users and the role of English as a trendsetter in language use in digital communication, contribute to developments such as the creation and rapid spreading of neologisms all over the Spanish speaking Twitter community, as in *covidiota*. The prominent role of English in this context

does not, however, keep English-induced items from being perceived as a conspicuous, unfamiliar, potentially marked, and, in any case, generally new item. Speakers typically have several strategies at their disposal for dealing with such linguistic items. The following section provides a short overview and possible classification of such strategies.

Dealing With New Borrowings: Alterity Markers

In the initial phase of the introduction of a borrowing into a receiving language, the new item is generally perceived as foreign and unfamiliar by speakers of that language. The fact that it does not (yet) belong to the vocabulary inventory of the receiving language often becomes manifest in the form of certain strategies which speakers apply to deal with the newness and foreignness of the newly borrowed item.

Following the works of Rey-Debove (1978, 1998) and Authier-Revuz (1995), Pflanz (2014) establishes a typology of "marqueurs d'altérité" ('alterity markers,' Winter-Froemel, 2021). These alterity markers are used in the context of different linguistic strategies applied by speakers to mark borrowings which are still considered to be new and unfamiliar, and therefore not taken to be self-explanatory (Pflanz, 2014, p. 164–165). Pflanz introduces four types of alterity markers. The first type (X montré, Pflanz, 2014, p. 164) comprises visual markers which make the item stand out in a text. For example, this can be achieved by the use of quotation marks. The second type (X commenté, Pflanz, 2014, p. 165–168) refers to metalinguistic comments which are made by the speaker and can either be of an evaluative nature (commentaire axiologique) or rather of an explanatory kind (commentaire didactique). The third type of alterity-marking strategy consists in paraphrasing, or translating, the new item to clarify its meaning (X traduit, Pflanz, 2014, p. 168–173). The fourth type, on the other hand, comprises the explanation of the neologism (X expliqué, Pflanz, 2014, p. 173–175). The fourth type differs from the third mainly in the predominant function associated with it, which is the referential one, while for X traduit the metalinguistic function is predominant (Pflanz, 2014, p. 173).

These categories or types, established by Pflanz, are taken into account at different points throughout the present study and are reflected in several of the research questions outlined in the following section. Since these features are explicitly established as markers of alterity, i.e., otherness, one very interesting question is whether the time span of the observation periods in the case study is long enough for their manifestations to change, as speakers become more familiar with the term at hand.

THE INTRODUCTION OF COVIDIOTA INTO SPANISH: RESEARCH QUESTIONS AND AIM

The subject of the present case study is the word covidiota, which was first created as covidiot < COVID + idiot in English during the early weeks of the coronavirus-crisis in the late winter of 2020. Its meaning can very broadly be described as 'someone who behaves in an inappropriate way with regard to the situation caused by the outbreak of the virus,' and, more specifically, was

initially used to refer to people who refused to adhere to the new rules and instructions pronounced by the authorities to curb the spreading of the virus. The new term spread quickly on several social media platforms, became widely known and was adopted by a lot of other languages, too (e.g., Italian and Portuguese: covidiota, German: Covidiot), with minor changes in form and/or grammatical properties (as for example the possibility to be used as a noun or an adjective in Spanish, while the English term was coined as a noun).

The creation of neologisms during, and in the context of, the COVID-19 pandemic has become a subject of interest, and the first publications dealing with COVID-induced creation of new words appeared not long after the beginning of lockdowns and the proclamation of a state of emergency in many countries all over the world. For example, several collections of corona-neologisms (those that emerged in the Spanish language, but also items from other languages) appeared in online newspaper articles. Apart from Engl. covidiot and Span. covidiota, there are other examples of potentially humorous blends which appear in these collections. Among them are other English-induced corona-neologisms, such as, for example, covidivorcio (e.g., Clarín, 2020a), which, just as covidiota, can be considered a borrowing (< Engl. COVID-19 + Engl. divorce) with an adaptation to Spanish noun morphology (divorce > divorcio), or a creation from Engl. COVID-19 + Span. divorcio (a loanblend, cf. Roig-Marín, 2020, p. 1; see below), constructed in Spanish following the English model covidivorce. The blend zoompleaños < Zoom + Span. cumpleaños (e.g., Clarín, 2020b) also constitutes a case of an English-induced blend, albeit a slightly different one. In this case, the first constituent, Zoom, is a proper noun (i.e., the name of a US-company and software nowadays widely known and used to carry out video conferences) and the second constituent is a Spanish element. Moreover, there are creations built from constituents which originate from other languages but have become lexicalized in Spanish, e.g., balconazi < balcón (< Ital. balcone, according to DLE, 2021, s.v. balcón) + nazi (< Ger. Nazi): 'those who fancy themselves balcony-policemen in order to insult those who are out and about in the streets' ("[l]os que ejercen de policías de balcones para insultar a quienes circulan por la calle," Pons Rodríguez, 2020). And there are COVID-related blends which do not per se display any foreign influence, as is the case with cuarenpena < Span. cuarentena + Span. pena (Pons Rodríguez, 2020).

Apart from newspaper articles, corona-neologisms appear, for example, in different collections of neologisms (which do not focus exclusively on COVID-related innovations but include them in their inventories), such as Antenario, provided by Antenas Neológicas—Red de neología del español (Adelstein and Freixa, 2021). The Antenario does not list covidiota, but includes, among others, covidivorcio and zoompleaños.

COVID-induced lexical innovation has also become subject to a considerable number of studies discussing the matter from a linguistic point of view. The neologism Span. covidiota can be found in some of them. Rivas Carmona and Calero Vaquera investigate memes which spread in WhatsApp chats in Spain between March 14 and June 21, i.e., during the state of alert (2020, p. 111). The files containing definitions of the neologism covidiot(a), either in English (see also Section From Defining

FIGURE 1 | English definition of *covidiot* shared on Twitter (https://pbs.twimg.com/media/ETqO70GWsAEEX9A?format=jpg&name=medium).

FIGURE 2 | Spanish definition of *covidiota* shared on Twitter (https://pbs.twimg.com/media/ETuYeEdXQAEId52?format=jpg&name=900x900).

to Referencing—Examples and Criteria, **Figure 1**) or in Spanish (Rivas Carmona and Calero Vaquera's example shows a version similar to, but not identical with **Figure 2**)[1] are categorized as a meme of textual content, and subsumed under the group of memes categorized as "[m]emes que entrañan comportamientos lúdicos" ('memes that convey a ludic conduct', Rivas Carmona and Calero Vaquera, 2020, p. 126). The authors emphasize the creative and playful character of the lexical innovation and the humorous effect it aims at, as well as the fact that the definitions use expressions clearly identifiable as negative, or derogatory, to describe the meanings of *covidiot(a)* (Rivas Carmona and Calero Vaquera, 2020, p. 126).

Rodríguez Abella also mentions *covidiota* among the newly created words she examines (2021, p. 81). Roig-Marín points out two predominant initiating centers of creation of new words in the COVID-19 pandemic: journalism and the social media platform Twitter (2020, p. 1). She refers to Span. *covidiota*, (or, rather, the plural form *covidiotas*) as a *loanblend*, with the second part of the blend appearing "in its Spanish rendering" (Roig-Marín, 2020, p. 1).

Ladilova, in her survey of word formation in Spanish in relation to the coronavirus-crisis, underlines the specific relevance of word formation mechanisms otherwise often considered to be rather marginal phenomena: "[…] 'minor' word formation processes such as abbreviations and acronyms are particularly productive in the context of the COVID-19 pandemic" (2020, p. 44). As she subsumes blending (Ger. *Kontamination* or *Amalgamierung*) under the general process of abbreviation (in English also referred to as clipping), this also includes the formation of *covidiota*, even though Ladilova points out the alternative possibility to classify *covidiota* as a compound, the merging of the syllable *-id-* then being caused

by morphophonemic reasons (Ladilova, 2020, p. 55). Rodríguez-Ponga documents and presents 127 items of Spanish COVID-related vocabulary, which, apart from *covidiota*, comprise the term *covidiccionario* "la recopilación del léxico covídico" (2020, p. 221). As for *covidiota*, the author describes its word formation as a case of composition from *covid* + *idiota*, involving a process of haplology which accounts for the loss of the syllable *-id-* in the middle of the compound (Rodríguez-Ponga, 2020, p. 221).

COVID-related neologisms have of course also been coined in and adopted by many other languages than English and Spanish. For German, for instance, the IDS (Leibniz-Institut für Deutsche Sprache) provide documentation of new, corona-induced vocabulary in the form of an online dictionary (*OWID – Neuer Wortschatz rund um die Coronapandemie*, Klosa-Kückelhaus et al., 2020b[2]). An early collection of COVID-related lexical items was published in a print version in May 2020 (Klosa-Kückelhaus et al., 2020a) and contains, among others, Ger. *Covidiot*. Apart from the online-dictionary, a series of *Stellungnahmen zur Sprache der Coronakrise* ('comments on language in the coronavirus pandemic'), i.e., articles dealing with the influence of the pandemic on the German language, have been published by the IDS (Klosa-Kückelhaus, 2020a,b; Möhrs, 2020; to give but a few examples).

The Spanish language academy, Real Academia Española (RAE), included the term *covidiota* in its historical dictionary *Diccionario histórico de la lengua española (DHLE)* in March 2021. There, it is depicted as being both a noun and an adjective, and categorized as "[c]alco estructural del inglés *covidiot*" (*DHLE*, 2013, s.v. *covidiota*), i.e., a 'structural calque from Engl. *covidiot*.' The *DHLE*-entry lists three meanings for *covidiota*:

---

[1]The main difference between the meme cited by Rivas Carmona and Calero Vaquera (2020) and **Figure 2** is that the former includes a fictitious explanatory note about the etymology of the term, saying "Del lat. covidiōta, y este del griego […] covidiōtēs" (2020, p. 126), "From Latin covidiōta, and that from Greek covidiōtēs".

[2]The resource can be accessed here: https://www.owid.de/docs/neo/listen/corona.jsp (accessed October 20, 2021).

1. *adj.* [Persona] Que se niega a cumplir las normas sanitarias dictadas para evitar el contagio de la covid.
2. *s. m. y f.* Persona que se niega a cumplir las normas sanitarias dictadas para evitar el contagio de la covid.
3. *adj.* Típico o característico de un covidiota.
   (*DHLE*, 2013, s.v. *covidiota*)

'1. *adj.* [Person] Who refuses to comply with the sanitary norms issued in order to avoid the proliferation of covid.
2. *noun*, *masc.* and *fem.* A person who refuses to comply with the sanitary norms issued in order to avoid the proliferation of covid.
3. *adj.* Typical or characteristic of a covidiot.'

Apart from these definitions, the entry comprises examples of usage for the different meanings, which originate from (online) newspaper articles from different countries. The *DHLE*-entry, moreover, offers the possibility to display related words (*familia de palabras*). In the case of *covidiota*, two derivatives are shown: *covidiotismo* and *covidiotez*; moreover, the etymology of the term is illustrated by linking Span. *covidiota* to Engl. *covidiot*, and the latter to Engl. *idiot* as well as Engl. *covid*, which, in turn, is linked to Engl. *covidivorce* (*DHLE*, 2013, s.v. *covidiota*—Familia de palabras).

The present paper focuses on the question of how *covidiot(a)* is introduced into Spanish. The study is conducted on the basis of a corpus of tweets in Spanish, which are analyzed in order to gain insight into the various relevant aspects of its way into the receiving language. The central questions can thus be formulated as follows: Is the newness, or otherness, of the term *covidiota* signaled or reflected in the tweets, and what indicators of foreignness can be found? What can be said about processes involving an increasing integration of the item into the Spanish language: do signs of alienness become less frequent over time? In order to approach the central aim of the study, several subordinate research questions are developed to investigate relevant aspects or sub-processes of the introduction of the term *covidiota* into Spanish.

Signs of Newness

The first relevant aspect to be outlined is concerned with form rather than with the content of the tweets: Speakers (and, of course, writers) often feel the urge to highlight items in their texts which are considered unusual as they are not part of the general vocabulary of their language because they are of foreign origin, for example. In the graphic medium (which includes, among others, the communication on social media platforms), such strategies comprise capitalization or the use of quotation marks, among others. The highlighting of foreign items corresponds to the first type of alterity markers as established by Pflanz (2014, p. 164–165): visual markers of alterity. The working hypothesis for the case study is that the use of these markers becomes less frequent over time, as the term becomes more widely known and continually loses part of its alienness.

At the same time, graphical highlighting is not only relevant when it comes to flagging recent borrowings, but it is also considered to be one of many different possible humor- or irony-markers available to speakers/writers who (consciously

or unconsciously) wish to underline an ironic or humoristic intention of their utterances ("typographic markers," Burgers and van Mulken, 2017, p. 388). As has been outlined above, blends are typically associated with playful language use and humoristic intention. *Covidiota* can therefore be considered at least a potentially humorous lexical item. So, if graphical highlighting should prove to be a prominent feature during all three observation periods, this feature would not (only) be used as a marker of newness but as a means to emphasize the speaker's humoristic intention.

A particular kind of graphical highlighting has to be discussed in more detail, since it is a highly characteristic feature of language in social media: the use of hashtags. Apart from highlighting in itself, the appearance of a term in the form of a hashtag often goes hand in hand with that term not being integrated into the syntactic structure of the rest of the tweet. Therefore, the appearance of *covidiota* or *covidiot* as a hashtag constitutes one of the categories of analysis.

When it comes to examining the way in which English, or, more precisely, the presence of an English-induced neologism such as *covidiot(a)* influences Spanish twitter discourse, it seems appropriate to look for other English (and English-induced) elements (words, phrases, indications of code-switching) in these texts. The underlying research question connected with this category of analysis is whether the presence of Engl. *covidiot* or Span. *covidiota* might trigger the use of more English-induced linguistic items.

Integration of *covidiota* Into Spanish Language Usage on Twitter

After establishing parameters that reflect the newness of the term *covidiot(a)* to the Spanish language, I will now propose a number of features that can be considered indicators of integration. These can be found on different levels of language and can be said to reflect a growing tendency of the term to behave as a Spanish word.

Although the neologism Span. *covidiota* is of English origin, its formation is completely transparent to Spanish speakers and it does not appear absolutely clear whether it is more appropriate to say that Spanish speakers (i.e., Spanish speaking Twitter users) borrow the entire new-built term *covidot* from English and then adopt it, or that they rather form the Spanish equivalent of the term, following the English example. The latter would be just as plausible as the first possibility, since both constituents *COVID* and *idiot(a)* can be considered internationalisms, albeit of very different kinds. While *COVID* is a very recent creation and was coined in English, Span. *idiota* (as well as Engl. *idiot*) is of ancient Greek origin and constitutes a long-established element of the Spanish lexicon.

Both the English and the Spanish versions of the neologism can be described as *cruces subordinativos* 'subordinating blends,' which means that one of the two components constitutes the nucleus of the word (Ulašin, 2016, p. 165). In this case, it is the second constituent: a *covidiot(a)* is a particular kind of *idiot(a)*.

The nucleus is also decisive when it comes to determining the grammatical properties of the blend, for example its belonging

to a certain word-class. The English term *idiot*, according to the *OED*, can be used as a noun or an adjective (*OED* online, 2021, s.v. *idiot*), although the adjective *idiotic* appears to be the more frequently used alternative. The definitions of Engl. *covidiot* which are received, adopted and disseminated by Spanish speaking Twitter users during the initial phase of its existence, characterize it as a noun. As for the Spanish term *idiota*, the Spanish language academy, (RAE), in its *Diccionario de la Lengua Española* (DLE), classifies it as an adjective, but points out the fact that it is also used as a noun, by adding the annotation "U[sado] t[ambién] c[omo] s[ustantivo]" (*DLE*, 2014: s.v. *idiota*). Other dictionaries register *idiota* as both an adjective and a noun (e.g., Moliner, 2008: s.v. *idiota*).

The actual behavior of Span. *covidiota* with regard to word-class will therefore be one of the aspects under investigation, the actual question being: If the English etymon *covidiot* has been coined as a noun, and Span. *idiota* can be an adjective as well as a noun, how does this influence the use of *covidiota* over time? Is it introduced into Spanish as a noun and continually used (mostly) as such, or is there an increase in usage as an adjective?

When dealing with a foreign, unknown lexical item, it can be expected that users feel a need to define it and comment on the new term before they start actually using it. This point relates to the second, third and fourth types of alterity markers as established by Pflanz (2014) and, with regard to the study at hand, leads to the following crucial questions: Is the term *covidiota* talked about, cited and/or explained, or is it actually used? Does the main function seem to be referential or rather metalinguistic? How does the use of *covidiota* change over time, with regard to these aspects? The differentiation between metalinguistic commenting, defining, negotiation of semantic content and referencing is thus considered a crucial parameter for the analysis of the tweets. With regard to the above-mentioned typology, a possible hypothesis could be that commenting (second type), translating (third type) and explaining (fourth type) become less frequent functions over time, since all three are considered as alterity markers and thus descend in frequency as the item in question becomes better-known.

Another indicator which is considered highly revealing in terms of integration of a new item into its receiving language is the formation of new phrases, compounds, etc. which include the neologism in question. Therefore, one important point of investigation in the present case study consists in the search for recurring expressions in Spanish tweets which contain the term *covidiota*, the hypothesis being that such phrases should become more frequent as the term becomes more widely known and more frequently used.

MATERIALS AND METHODS

As the case study is intended to focus on a qualitative examination of the outlined questions, the number of tweets to be analyzed had to be limited. Therefore, the material subjected to analysis in the study consists of tweets from three separate weeks in 2020, starting with the week in which *covidiota* could be found for the first time in a tweet in Spanish. The second and third

weeks under investigation follow with a distance of 14 weeks from the previous one.

The tweets were retrieved manually, with the help of Twitter's advanced search function, applying the following search parameters:

"All of these words"   *covidiota*
"Language"   Spanish
"Dates: from … to …"   e.g., from 29 June 2020 to 30 June 2020

The tweets were then anonymized, deleting any names and references to private Twitter users (including names of cities, etc., which might reveal information concerning a user's identity)[3], with the exception of public figures such as widely known politicians (heads of states, ministers of governments, etc.), well-known celebrities and institutions. The anonymized tweets were subsequently subjected to analysis along several different aspects. The following properties were established as parameters for the analysis:

1) graphic representation (focus on means of graphical highlighting)
2) presence of #: use as a hashtag or without it
3) language: English form (*covidiot*) vs. Spanish form (*covidiota*)
4) other English items (words, phrases, etc.)
5) parts of speech: *covidiota* as adjective or noun
6) *covidiota* in the Spanish sentence: integration vs. unconnected use
7) metalinguistic use vs. definition vs. referencing
8) recurring constructions, patterns of combined occurrence
9) major shifts in meaning

These parameters are considered relevant indicators with regard to the central research aim as stated above. Although the study is of a qualitative nature, the results regarding the individual research parameters are presented in numbers (usually both absolute and relative), so as to make it possible to spot tendencies and developments over time. It has to be underlined, however, that the numbers do not claim any statistical significance. The small sample size and selective nature of the observation periods do not allow for such claims.

The examples cited below are quoted exactly as they appear in the corpus. Spelling mistakes, ungrammatical expressions, etc. are neither pointed out as such, nor are they corrected or changed. The translation of emoticons into their corresponding verbal description is maintained. A typographic distinction is made between the actual tweet text (in italics) and replacements (anonymizations) as well as paraphrased emojis (in regular letters).

RESULTS

The number of results, after the removal of false positives, i.e., tweets in other languages than Spanish (Italian, Portuguese, and Catalan), for the three periods of time (T1–T3) indicated in

---

[3]The names of people were therefore replaced by <NAME> in the corpus, the names of cities by <CITY>, etc.

TABLE 1 | Number of results for *covidiota* in Spanish tweets, for the indicated time periods.

Time period	Dates	Number of tweets	Tokens for Span. *covidiota(s)*	Tokens for Engl. *covidiot(s)*
T1	16–22 March 2020	136 tweets	138 tokens	20 tokens
T2	29 June–5 July 2020	650 tweets	649 tokens	18 tokens
T3	12–18 October 2020	705 tweets	731 tokens	2 tokens

section The Introduction of *covidiota* Into Spanish: Research Questions and Aim, is represented in **Table 1**.

The word form entered as a search term in Twitter's search mask was *covidiota*. Nonetheless, in several cases the search returned tweets which contained not this exact form, but the English form *covidiot*. These results were not removed but included in the analysis, provided the tweet-text surrounding the target term was basically in Spanish (allowing, of course, for other contact-induced elements appearing alongside the target term). In T2, for example, 650 tweets comprise 649 tokens for the Spanish form plus 18 tokens for the English form, which means that at least one tweet does not actually contain the Spanish term, but uses the English term while the surrounding text is composed in Spanish.

As can be seen in **Table 1**, the number of tweets which contain the item in question is rising over time, the difference between the first and second period being considerably larger than that between T2 and T3. At the same time, the number of tokens for the "original" English term, *covidiot* (including both the use as a hashtag and without one) descends considerably from T1 to T3. The following sections are going to provide more detailed information about the results regarding the individual research parameters set up in the previous chapter.

Signs of Newness: Results
Graphical Highlighting

Graphical highlighting is considered a relevant means to convey a notion of newness of a linguistic item, as laid out above. Differences can be made out with regard to the scope as well as to the particular kinds of highlighting. It can include an entire phrase or sentence, or just the word itself, or even only parts of the word. As for the different ways of highlighting, two basic types can be distinguished: capitalization, on the one hand, and the use of quotation marks on the other. **Table 2** shows the results according to units and means of graphical highlighting.

As **Table 2** shows, the proportion of tokens including any type of graphical highlighting descends by a large degree from T1 to T2. The difference between T2 and T3, however, is rather small. It can thus be said that the material analyzed for this study shows a marked decline in the frequency of visual alterity markers, but only with regard to the first two periods under investigation. The further development from T2 to T3 suggests a certain continuity with regard to this particular parameter, at least for the second half of the time span under investigation.

English Variant and the Use of Hashtags

Although the search parameters chosen for the retrieval of samples for the corpus explicitly included the Spanish form *covidiota*, the search returned several results which contained the English etymon, *covidiot*, while the overall language of the tweet was Spanish. These tweets were not sorted out but included in the corpus, since the mere fact of their existence seems significant for the research aim at hand. When investigating the introduction of a new anglicism into the Spanish language, this can be considered an essential stage of the process.

As **Table 3** shows, the appearance of Engl. *covidiot* is considerably more frequent when it functions as a hashtag than without one. Moreover, both *#covidiot* and *#covidiota* can be found in all three observation periods, albeit with decreasing frequency. Decreasing frequency can also be observed for the use of the English term *covidiot* in the Spanish tweets examined in the course of this study, in both forms [i.e., *covidiot* and *#covidiot(s)*].

The use of hashtags, as has been pointed out before, constitutes a special type of highlighting in its own right. In terms of integration into the sentence structure, however, both hashtags, the English and the Spanish form, can appear as part of the sentence as well as in a peripheral, more or less unconnected position.

(1) *El presidente de* COUNTRY *recomienda protegernos del coronavirus con estampitas y amuletos #covidiota* [98]
  'The president of COUNTRY recommends that we protect ourselves from the coronavirus with pictures of saints and amulets *#covidiota*'

(2) *Sera muy #covidiota de mi parte entrar a clases de natación con* NAME? *Necesitamos distraernos*
   Weary face [578]
  'Would it be very *#covidiota* of me to attend swimming classes with NAME? We need to take our minds off things'
   Weary face

(3) *Sigue mucho #Covidiot en las calles de* CITY [583]
  'There are still many *#Covidiot* in the streets of CITY'

In (1), the hashtag is placed at the end of the tweet and not integrated into the sentence structure. There are, however, also cases where they function as fully integrated parts of the sentence, as can be seen in (2) and (3).

Other English or English-Induced Elements in the Tweets

As **Table 4** shows, 3.68% of all the tweets in T1, 8.92% of those in T2 and 10.35% in T3 contain English or English-induced items. Not included are proper nouns, for example names of cities or locations. The terms *Facebook*, *Twitter,* and *Instagram* were excluded from the count as long as they were used as names (to refer to the respective social media platforms), but were included as soon as they appeared as a constituent of a different word class (i.e., not a name), such as *twittear* or *tuitear*, for example.

For T3, a separate line shows the number of tweets which contain the word *Halloween* (including several spelling variants). These tweets were counted as a separate group, since their appearance is clearly linked to the season in which the samples

TABLE 2 | Results for graphical highlighting.

Time period		T1	T2	T3
Total number of tokens (*covidiota* + *covidiot*)		**158**	**667**	**733**
Capital letters: phrase/sentence	e.g., *NO SEAS COVIDIOTA*	2	5	4
Capital letters: (Spanish) word only	*COVIDIOTA*	25	19	15
Capital letters: English word	*COVODIOT*	12	9	2
Capital letters: part of the word	e.g., *covIDIOT, covIdiot, COVIDiot,* etc.	7	10	19
Capitalization of the first letter[a]	*Covidiot*	24	41	55
Quotation marks[b]: phrase/sentence	e.g., *"no seas covidiota"*	7	5	1
Quotation marks: word only	*"covidiota"*	8	10	4
Others	e.g., **no seas covidiota*, *<covidiota>,* etc.	1	2	2
TOTAL (absolute)		86	101	102
TOTAL (relative)		54.43%	15.14%	13.92%

[a]*This feature is not registered as a means of highlighting if it appeared at the beginning of a tweet or sentence, or after punctuation characters generally used to mark the end of a sentence.*

[b]*This category includes all common variants of quotation marks, i.e. "_" as well as '_' and ≪_≫. It is possible that the function of single quotation marks, even in their use in social media, differs slightly from that of the others. However, this possible difference is not taken into account in the present study.*

TABLE 3 | *Covidiot,* #*covidiot(s),* and #*covidiota(s)*: numbers per observation period.

	in T1 (abs.)	in T1 (rel.)	in T2 (abs.)	in T2 (rel.)	in T3 (abs.)	in T3 (rel.)
covidiot	4	2.94%	2	0.31%	0	0%
#*covidiot(s)*	15	11.03%	17	2.62%	2	0.28%
#*covidiota(s)*	40	29.41%	10	1.54%	31	4.40%

TABLE 4 | English and English-induced elements in the tweets.

	in T1 (abs.)	in T1 (rel.)	in T2 (abs.)	in T2 (rel.)	in T3 (abs.)	in T3 (rel.)
English items	5	3.68%	58	8.92%	73	10.35%
Special case: *Halloween,* incl. variants (*Jalogüin, jalowin,* etc.)					39	5.53%

were retrieved. Moreover, although the term is also English-induced, it has to be regarded as a name rather than a common noun and therefore should not be included in the same category as the others.

The presence of English or English-induced items in the tweets, apart from *covidiot(a)*, is a recurring feature within the analyzed tweets, the highest proportion being shown for T3 with 10.35%, and the tendency in this case appearing to be rather an increasing, not a declining one. As the results suggest that, at most, about every 10th of the examined tweets contain other linguistic items in English, they do not appear to confirm the hypothesis that the presence of *covidiota* might trigger an intensified use of anglicisms. Neither do they clearly reject it. In order to provide any definite conclusions of the results in this respect, a comparative study would have to be carried out to provide information on the proportion of anglicisms in tweets produced by a similar sample of twitter users, so as to see whether that proportion is similar or not.

The English or English-induced items considered in the analysis are, for the purposes of the present study, considered as one single group, even though there are actually different kinds of items contained within that group. In the first place, they are not very numerous in the tweet-corpus at hand. Secondly, what matters with regard to the research aim is the presence or absence of other English-induced items. Distinguishing different types of anglicisms, on the other hand, does not constitute a matter of high priority in this respect. Nonetheless, some examples shall be given to illustrate some individual cases:

(4) *Todos diciendome cosas de que mi roomie es covidiota en vez de decirle algo directamente a ella*
Face vomiting [687]
'Everyone telling me things about my *roomie* being *covidiota* instead of saying something directly to her'
Face vomiting

(5) *Estoy a nada de ser covidiota, ya no awanto el encierro jelp* [757]
'I'm on the verge of being *covidiota*, I can no longer stand being locked up *jelp*'

(6) *estoy a nada de ser una covidiota e ir a estudiar a un starbucks, i cant stay in my room no more* [987]
'I'm on the verge of being a *covidiota* and going to a Starbucks to study, *i can't stay in my room no more*'

TABLE 5 | Parts of speech – *covidiota* as adjective or noun in the corpus.

Observation period		Tokens (total)	*covidiota* as adjective (tokens)	*covidiota* as noun (tokens)	*covidiota* as adj. or noun: indeterminable (tokens)
T1	abs.	138	10	58	70
	rel.	100%	7.25%	42.03%	50.72%
T2	abs.	649	218	321	110
	rel.	100%	33.59%	49.46%	16.95%
T3	abs.	731	386	270	75
	rel.	100%	52.80%	36.94%	10.26%

TABLE 6 | Isolation of *covidiota* in the Spanish sentence.

Observation period	*Covidiota* as only text (one-word-tweets)	Isolation/segmentation			Non-integrated use of *covidiota*: Total
		Covidiota in initial position	*Covidiota* in final position	Others	
T1 abs.	11	31	31	2	75
T1 rel. (136 tweets)	8.09%	22.79%	22.79%	1.47%	55.15%
T2 abs.	36	6	16	1	59
T2 rel. (650 tweets)	5.54%	0.92%	2.46%	0.15%	9.08%
T3 abs.	9	1	9	1	20
T3 rel. (705 tweets)	1.28%	0.14%	1.28%	0.14%	2.84%

(7) *Hoy traigo mucho mucho hate para los vale verga #covidiota #COVID19* [1305]

'Today, I'm carrying a lot of *hate* for those who just don't care #covidiota #COVID19'

While in (4) and (7) individual lexical items (*roomie* and *hate*, respectively) are inserted into the Spanish text without showing any visible alteration in form, in (5) English *help* has changed into *jelp*, probably representing a pronunciation variant which replaces the sound [h], which is not part of the phoneme inventory of Spanish, by [x]. Apart from that, (5) is another example of a single lexical item being included in a Spanish tweet text. In (6), however, the English item is not one individual word, but an entire sentence: *i* [sic] *can't stay in my room no more.*

Integration of *covidiota* Into Spanish Language Usage on Twitter
Parts of Speech: *covidiota* as an Adjective and/or a Noun

In the definitions circulating within social media platforms during the period of origin of the term, Engl. *covidiot* is usually classified as a noun. The same applies for most of the first definitions in Spanish which make an explicit reference to word-class. However, this does not necessarily mean that the Spanish term is really used as a noun, or, at least, it is not only a noun but also an adjective.

The representation in **Table 5** considers only the tokens for Span. *covidiota*, leaving out those for Engl. *covidiot*. The distinction between *covidiota* as a noun and *covidiota* as an adjective is not always very clear and, in some cases, cannot be made at all. This is mostly the case if *covidiota* appears without a clarifying co-text from which to gather this information [e.g., in one-word-tweets, but also in examples like (8)].

(8) *Ya acuñaron una nueva palabra, "covidiota".* [3]

'They've just coined a new word, "*covidiota*"'.

It can be seen in the last column of **Table 5** that the proportion of such unclear cases decreases over time. This development can be attributed to the fact that the term is increasingly used and integrated into a Spanish sentence structure, rather than being cited, defined, or talked about, as is often the case in the first phase of its existence.

Covidiota in the Spanish Sentence: Integration vs. Isolation

Table 6 shows the number of cases (tweets) in which *covidiota* is not integrated into the Spanish sentence structure but stands apart, occupying, in one way or another, a marginal position within the tweet. In these cases, *covidiota* most frequently appears either in an initial position [often followed by a definition, as in (9)], or in a final position [often as a hashtag, as in (10)].

TABLE 7 | Metalinguistic use, defining, referencing and combination of functions.

Observation period: abs./rel. (total of 1.491 tweets)	T1 abs. (136)	T1 rel.	T2 abs. (650)	T2 rel.	T3 abs. (705)	T3 rel.
Metalinguistic commenting only	9	6.62%	14	2.15%	2	0.28%
Definition only	23	16.91%	3	0.46%	1	0.14%
Metalinguistic commenting + definition	33	24.26%	2	0.31%	1	0.14%
Unconnected mentioning of the term	7	5.15%	10	1.54%	5	0.71%
Referencing + definition (separated)	13	9.56%	0	0%	0	0%
Referencing only	37	27.21%	444	68.31%	381	54.04%
Referencing and negotiating semantic content	8	5.88%	153	23.54%	307	43.55%
Containing a metalinguistic comment (and, possibly, other functions)	45	33.09%	19	2.92%	8	1.13%
Containing a definition (and, possibly, other functions)	74	54.41%	6	0.92%	1	0.14%
Containing referencing	63	46.32%	602	92.62%	693	98.30%
Containing use of *covidiota* as/in a name	0	0%	21	3.23%	3	0.43%

(9) *Covidiota*
    Dicese de la persona que no respeta la cuarentena [41]
    '*Covidiota*
    It is said about the person who doesn't adhere to quarantine'

(10) (=1) *El presidente de* COUNTRY *recomienda protegernos del coronavirus con estampitas y amuletos #covidiota* [98]
    'The president of COUNTRY recommends that we protect ourselves from the coronavirus with pictures of saints and amulets *#covidiota*'

One-word-tweets, where the word *covidiota* actually constitutes the only written text present, represent a special case of isolated use and are therefore also included. It has to be noted, however, that a one-word-tweet can be part of a thread or series of reciprocal communication between two or more Twitter users, so that it does not necessarily have to be entirely isolated when taking into account the expanded co-text.

The numbers in **Table 6** show an overall tendency toward a growing integration of the term into the Spanish sentence structure. While in T1 more than half of all the tweets containing the word *covidiota* do not include it in their syntactic structures, in T3 this proportion has dropped to <3%. This development can be closely linked to another one: the changes of proportion with regard to metalinguistic use vs. use for referencing, which will be presented in the following section.

Metalinguistic Use and Defining vs. Referencing (Integration in Terms of Content)

In the first observation period, a majority of tweets contains a definition of the term Engl. *covidiot* and/or of Span. *covidiota*. As for the allocation of the tweets to the categories indicated in **Table 7**, it has to be noted that there are numerable cases which apply to several categories (for examples and further aspects of distinction between and combination of categories, see section Discussion).

When *covidiota* is used as an isolated item (i.e., the tweet consists of just that one word), one possibility is that it represents a direct response to a preceding tweet, e.g., a picture which shows a group of people who do not wear their masks as recommended. In these cases, *covidiota* in the following tweet is interpreted as used by the author to refer to the people in the picture (to condemn the behavior shown in the preceding tweet). In other cases, a tweet containing only the one word, *covidiota*, can be accompanied by an image file, showing a definition (either in English or in Spanish) of the term. These cases then belong to the category "definition." There are, however, tweets for which neither of the two options applies, i.e., they do not, in any ascertainable way, respond to a preceding tweet, and neither do they introduce or accompany a definition of the term *covidiot(a)*. These cases therefore constitute their own category of "unconnected mentioning," since no context can be identified.

Recurring Constructions, Patterns of Combined Occurrence

In the course of the analysis of the 1,491 tweets, some combinations of *covidiota* with other linguistic items could be made out as recurring patterns. One such combination is the combination of the copula *ser* 'to be' + *covidiota*. (11) is a further example for this type.

(11) *yo quiero ser covidiota*
    Happy man raising one hand [454]
    'I want to be *covidiota*'
    Happy man raising one hand

Another recurring expression is *andar de covidiota*, Engl. (roughly) 'to be out and about as a covidiot,' which appears 60 times within the corpus (T2: 17 times, T3: 43 times). With regard to similarly constructed expressions, the following expanded pattern can be made out: verb + *de* + *covidiota*. The

corpus provides evidence of this pattern for the verbs *seguir* 'to continue' (two cases), *salir* 'to go out' (three cases), *ir(se)* [a algún lugar] 'to go [somewhere]' (12 cases), and *estar* 'to be' (five cases). Examples (12), (13), and (14) illustrate this type of expression.

(12) *Que pinches ganas de andar de covidiota* [1141]
'How damn much I want to be out and about as a *covidiota*'

(13) *hoy me fui de covidiota a la playa, que les digo* [866]
'today I went to the beach as a *covidiota*, what do I tell you'

(14) *¿O sea que aunque me encierre toda la vida si mi hermana sigue de covidiota igual me voy a morir?* [671]
'That is to say that, even though I lock myself up forever, if my sister continues to be a *covidiota*, I'm still going to die?'

(15) *creo que es un buen momento para decir que soy covidiota de closet, no me odien pls* [326]
'I think this is a good moment to say that I'm a covert *covidiota*, don't hate me *pls*'

Apart from these expressions, which, altogether, provide ways to include Span. *covidiota* into a verb phrase, there is another recurring expression: the noun phrase *covidiota de closet*, Engl. (roughly) 'covert covidiot'. It is less frequent than the aforementioned verb phrases but, nonetheless, worth mentioning. It can be found 17 times in the corpus and is exemplified in (15).

Major Differences in Meaning

With regard to major shifts in meaning, one development can be reported. The vast majority of Twitter users employ the term *covidiota* in a playful way, which could be interpreted as an attempt of humorous and self-ironic reflection of the unpleasant, alarming, and stressful situation brought about by the outbreak of SARS-CoV-2 and the measures imposed to curb its spreading among the population. When used in this sense, *covidiota* seems to adopt rather a broad meaning, including roughly any activity which takes place outside one's home. It appears to apply to any activity that includes meeting friends, but also to going to the beach, traveling, or going to the hairdresser's [see examples (16), (17), and (18)].

(16) *Si este fin voy a la playa, pero voy a la playa secreta que no va nadie, me catalogo un covidiota?* [420]
'If I go to the beach this weekend, but I go to the secret beach where nobody goes, do I classify as a *covidiota*?'

(17) *Me urge un viaje aunque suene como un covidiota* [1087]
Pleading face
'I urgently need to travel, even though I might sound like a *covidiota*'
Pleading face

(18) *me súper urge cortarme el cabello pero no quiero ser una covidiota* [355]
'I super urgently need to have my hair cut, but I don't want to be a *covidiota*'

As the examples show, it is a noteworthy feature of the term *covidiota* in this sense that it is very frequently used to refer to oneself. This fact supports the hypothesis that, at least in

some cases, a noticeable proportion of self-irony is included in these statements.

There is a very different reading of the term *covidiota*, which does not at all show any traces of self-irony. Instead, it conveys a strong sense of irony, even sarcasm (understood as a particularly strong way to express a negative attitude through irony; cf. Colston, 2017, p. 241), but its aim is not the speaker (tweet-author) him/herself, but another person or group (on the role of aggression in verbal humor and opposing groups as targets of humor, cf., e.g., the overview in Attardo, 2020, p. 64–67). This particular reading of covidiota represents an extremely noteworthy semantic change, since the initial meaning (*covidiota* 'someone who does not respect/follow the antivirus-measures') has been turned into its opposite, with the term now referring to 'someone who (unjustifiably, in the eyes of the speaker) considers the coronavirus a serious threat and supports, follows, implements, or even creates measures for its containment'. This meaning of *covidiota* is much less frequent in the corpus at hand than the other one, and in several cases, the particular meaning only becomes clear after consulting the context (threads and exchange of tweets before and after the tweet in question). (19) and (20) provide two examples.

(19) *¿Obligar a los niños a usar bozal? El secretario de "educación" es un covidiota* [1120]
'Make children use a muzzle? The secretary of "education" is a *covidiota*'

(20) *El covidiota no pregunta, solo asiente a la tv.* [883]
'The *covidiota* doesn't ask questions, but just agrees with TV.'

In contrast to the initial meaning discussed above, this newer (its first appearance within the corpus is in T2) and opposite meaning does not seem to come with one central, common definition which is then shared between users and subsequently translated into Spanish. The tweets which can be attributed to this reading of *covidiota* only belong to two of the above-established categories: "referencing and negotiating semantic content" or "referencing only." However, due to the very small number of tweets representing this particular meaning of *covidiota*, it is hard to reach any reliable conclusions about its use, beyond what has already been said.

DISCUSSION

Covidiota and Word-Class

The question of what part of speech *covidiota* actually constitutes in a specific case of language usage cannot always be answered in a definite way. There are cases which do not allow for a clear allocation to either one of the possible word classes (i.e., noun or adjective) at all. Other instances provide certain indications which can support the analysis. In the following, some examples shall be outlined in more detail. The *Nueva gramática de la lengua Española* (*NGLE*), published by the Real Academia Española, is used as a reference in cases of doubt and to provide further information.

Covidiota is very frequently used together with the verb *ser* 'to be,' which can generally be followed by an adjective as well as by a noun. Examples (21) and (22) illustrate a significant difference

between two different possible constructions with *ser*: in (21), *covidiota* is accompanied by the indefinite article *un*, and thus clearly used as a noun. In (22), however, there is no determiner, nor are there any other indicators which would encourage an interpretation of *covidiota* as a noun.

(21) *'No seas un CovIdiota.'* [5]
       'Don't be a *CovIdiota*'
(22) *No seas "covidiota"* [63]
       'Don't be *"covidiota"* ['covidiotic']'

Following the descriptions in the *NGLE*, in the present study, in uses like (22), i.e., constructions of the type *ser + covidiota* (without insertion of determiners), *covidiota* is taken to be an adjective. The same applies for the construction *por + covidiota*, as for example in (23), since these phrases can be interpreted as being essentially a shortened form of *por ser covidiota* (cf. *NGLE*, 2009, p. 3466), 'for being *covidiota* ['covidiotic']'.

(23) *si sales positivo es por covidiota* [208]
       'if you turn out to be positive, it's because of being *covidiota*'

The identification of word-class in constructions of the type verb *+ de + covidiota* (*andar/ir/estar/seguir de covidiota*, see Section Recurring Constructions, Patterns of Combined Occurrence) brings about certain difficulties. Here, *covidiota* is preceded by neither a determiner nor any other element which would allow for its identification as a noun at first glance. Neither is there any clear evidence supporting an interpretation of *covidiota* as an adjective. Consultation of the *NGLE*, however, suggests an interpretation of *covidiota*, in the instances indicated above, as *complementos predicativos nominales*.

> 38.10d No forman parte de complementos de régimen, pero se construyen con *de* […] los complementos predicativos nominales que designan alguna actividad u ocupación en la que se ingresa ([…] *Me voy a París de becario* […]), se permanece ([…] *Está de jefa de sección* […]) o se termina […]
> (*NGLE*, 2009, p. 2891).

The *complementos de régimen*, mentioned in the first part of the extract, are those whose preposition is determined by the verb, as would be the case for *de* in *quiero salir de aquí* 'I want to get out of here,' for example. This does not apply to the examples at hand, as in *andas de covidiota*, *sigue de covidiota*, etc. Moreover, the interpretation of the syntactic functions within the examples under investigation as copula constructions with predicative complements is highly plausible, even mandatory in the case of *estar de covidiota* (*estar* being one of the two central, prototypical copula verbs in Spanish, next to *ser*). This fact alone does not imply that the item following the preposition *de* necessarily has to be a noun, since there are also constructions of the type verb *+ de +* adjective, as in *acusar de incompetente* (*NGLE*, 2009: 2891) 'to accuse of being incompetent', also containing a predicative complement, albeit one that relates to the *complemento directo*. However, as is pointed out in § 38.10c, adjectival elements frequently appear in this kind of construction together with verbs that express judgement (*NGLE*, 2009: 2890–2891), in this case *acusar*. The verb semantics and the overall meaning of

the phrases from the *covidiota*-corpus (*andar de covidiota*, *ir de covidiota*, etc.) are of a different nature, and actually match the description in the extract cited above very closely: *designan alguna actividad u ocupación en la que se ingresa […], se permanece […] o se termina […]* – the activity or occupation in our case being that of 'acting as a covidiot,' 'going somewhere as/being a covidiot,' 'continuing to be a covidiot,' etc. Some of the examples from the extract also closely resemble some of the constructions in the corpus:

(24) *Creo que me voy a ir una semana de covidiota a* CITY *porque aún no quiero subirme a un avión.* […] [1214] ('I think I'll go to CITY as a *covidiota* for a week because I don't want to board a plane yet') compared to *Me voy a París de becario* ('I go to Paris as a scholarship holder'), and
(25) *que ganas de estar de covidiota en la playa* [959] 'How I would like to be at the beach as a *covidiota*' compared to *Está de jefa de sección* ('She works as head of department').

Thus, with reference to the RAE's assessment as quoted in the extract above, in expressions of the type *andar/ir/estar/seguir + de + covidiota*, for the part of speech-analysis at hand, *covidiota* is considered to be a noun. This decision suggests an interpretation following the RAE's description of the elements in question as *complementos predicativos nominales*, nominal predicative complements.

As a further example, I would briefly like to discuss the elements *muy* and *mucho* as accompanying items of *covidiota*. At first glance, both items seem to be rather unproblematic in terms of part of speech analysis and therefore can be of much help when it comes to determining the word-class of the element to be modified by them: *muy*, as an adverb, modifies the adjective *covidiota*, whereas *mucho* relates to *covidiota* as a noun (or rather, as *covidiota* has to be considered a countable noun, the correct form in standard Spanish should be the plural, as in *muchos* or *muchas covidiotas*). This holds true for *muy* in most cases, as is illustrated by the example (26).

(26) *Es que sí hay que ser muy covidiota para hacer una boda en estos tiempos*
       Man facepalming [355]
       'One really has to be very *covidiota* to have a wedding in times like these'
       Man facepalming
(27) *Mi hermana no tiene trabajo por culpa de covid, y la muy covidiota anda en el centro de compras jaja*
       Sneezing face Person facepalming [276]
       'My sister doesn't have a job due to covid, and the very *covidiota* goes out shopping in the mall'
       Sneezing face Person facepalming
(28) *Mucho covidiota el día de hoy*
       Thinking face [225]
       'A lot of *covidiota* today'
       Thinking face

Example (27), however, is much more ambiguous. On the one hand, the presence of the adverb *muy* would usually suggest a classification of *covidiota* as an adjective. On the other hand, *muy* is preceded by the definite article *la*, which provides a

strong clue for the interpretation of *covidiota* as a noun. In the *NGLE*, expressions similar to the one in question are mentioned in connection with a certain type of emphatic apposition, as in *la muy loca de tu prima* [...] (*NGLE*, 2009, p. 887). For that the authors refer to processes of neutralization of noun and adjective (*NGLE*, 2009, p. 887). Although they generally opt for the analysis as a nominalized adjective (cf. also § 13.7c, *NGLE*, 2009, p. 944), they explicitly point out the possibility of a reverse interpretation, i.e., analyzing the element in question as an adjective, or rather, an adjectival use of a noun, "ya que los nombres que se admiten en esta estructura suelen aceptar el adverbio *muy*, característico de los adjetivos graduables" (*NGLE*, 2009, p. 887). For the purposes of the present study, and following the *NGLE*, *covidiota* in expressions of this type (*la muy covidiota*) was analyzed as a noun, bearing in mind the fact that this is not the only possible way of interpretation.

The tweet-corpus does not contain the forms *muchos* or *muchas* in the plural. Instead, there are 10 instances of *mucho* in the singular, followed by *covidiota* (also in the singular), as illustrated in (28).

The use of the singular instead of the plural form does not alter the fact that *mucho* as a *cuantificador* helps identifying *covidiota* as a noun in these cases. The meaning, on a denotative level, is the same as that of *muchos covidiotas*. For the connotative level, however, the authors of the *NGLE* point out that, firstly, these constructions are more common in colloquial than in formal conversations, and secondly, that they often tend to convey a derogatory note (*NGLE*, 2009, p. 1479).

As these examples show, the analysis of Span. *covidiota* as a noun or as an adjective involves some particular challenges and cases of doubt. Bearing in mind the fact that uses are not always unambiguous and that decisions were made as outlined above, the numbers obtained in the present study and presented in **Table 5** suggest an increase in the proportions of the use of *covidiota* as an adjective over time.

When it comes to interpreting these results, the question is whether this growing tendency of *covidiota* to function as an adjective can be considered as a manifestation of increased integration into the Spanish language. This would be the case if the proportions of Span. *covidiota* as noun/as adjective could be shown to become more similar to those of Span. *idiota* as noun/as adjective, over time. Online corpora of Spanish were consulted in order to provide the necessary data to make this comparison.

The research in several online corpora of Spanish reveals a varied picture with regard to the use of Span. *idiota* as a noun or as an adjective. **Table 8** shows the results for three corpora of the *Corpus del español*, *Historical/Genre* (Davies, 2002), *Web/Dialects* (Davies, 2016) and *NOW* (*News on the web*, Davies, 2012), as well as for the RAE's *Corpus del Español del Siglo XXI* (Real Academia Española, 2021) and *Spanish Web 2018* (*esTenTen18*) in Sketch Engine (2021)[4].

As can be seen in the table, the proportions of *idiota* tagged as an adjective in the corpora range from 11.43% (*NOW*) to 51.02% (*CORPES XXI*), those for *idiot* as a noun from 48.96% (*CORPES*

XXI*) to 88.57% (*NOW*). The reasons for these considerable differences cannot be examined in detail in this study. However, it seems reasonable to assume that the classification is based on different criteria in the different corpora. In order to check, for instance, how *ser + idiota* is handled in the corpora, a search for the exact phrase *es idiota* was conducted, in order to test the results for this one specific form of *ser idiota*. In the *Genre/Historical*-corpus (Davies, 2002), there are nine cases of *es idiota* when looking for *idiota* as a noun, and none at all for *idiota* as an adjective. The *Web/Dialects*-corpus shows 520 hits for *es idiota* with *idiota* tagged as a noun, and 20 for *idiota* as an adjective, and the *NOW*-corpus contains 487 cases of *es idiota* with *idiota* as noun, but none with the adjective. In the *CORPES XXI*, on the other hand, all 17 instances containing the phrase *es idiota* are delivered when searching for *idiota* as an adjective, while non is returned when searching for the noun. Thus, there might be a general difference between the *Corpus del español* and the *CORPES XXI* regarding the handling of *ser idiota* in terms of word-class, which could constitute one of the reasons for the strongly varying numbers represented in **Table 8**.

A short examination of some selected samples reveals that sometimes, cases which appear to show the same structure appear in both groups, as a noun and an adjective. For example, the expression *el/la muy idiota* (similar to the above discussed *la muy covidiota*), is shown to appear 26 times in the *NOW*-corpus. In 24 of these cases, *idiota* is tagged as an adjective. In two instances, however, *idiota* is tagged as a noun. It does not become clear from the examples what causes these differences with regard to word-class. There are other expressions which show variation in this respect within several corpora (e.g., *por idiota* and *qué idiota* tend to appear in both groups), which cannot be examined in detail here. On the whole, this seems to concern rather a small number of examples, and might just be due to errors in the Pos-tagging. It is therefore rather unlikely to be the reason for the big differences in the numbers shown in **Table 8**.

In addition to the corpus research for Spanish *idiota*, a quick consultation of some corpora of English was carried out in order to gain some information about the use of Engl. *idiot* with regard to word-class. The search in the *COCA* (Davies, 2008), *GloWbE* (Davies, 2013) and *BNC* (Davies, 2004)[5] all lead to the same result in that none of them delivered any findings of *idiot* being used as an adjective. Of course, this does not necessarily mean that it actually is not used as an adjective, as other corpora might show evidence for its adjectival use. However, this use can probably be said to be less common than in Spanish, since in three different corpora, no evidence can be found.

The question underlying the parts-of-speech analysis was whether the introduction of *covidiota* into Spanish brings about an increase of usage of the term as an adjective. The data provides evidence for that. Moreover, the examination of corpus data was carried out in order to find out whether the usage of *covidiota* becomes more similar to that of *idiota* over time. A comparison of the numbers gained from the *covidiota*-data (**Table 5**) with the numbers in **Table 8** shows that the result for *covidiota* as an adjective in T1 (7.25%) is lower even than that

---

[4]The search in the latter one of the named corpora was executed via the function "Wordlist", filtering for adjectives and nouns, respectively, containing *idiota*.

[5]All accessed *via* https://www.english-corpora.org/ (accessed October 28, 2021).

TABLE 8 | Consultation of Spanish online corpora—*idiota* as noun vs. adjective.

Corpus	Corpus size	Span. *idiota* as a noun		Span. *idiota* as an adjective		Span. *idiota* total	
		abs.	rel.	abs.	rel.	abs.	rel.
Corpus del español –Genre/Historical	100 million words[a]	233	66.57%	117	33.43%	350	100%
Corpus del español –Web/Dialects	Two billion words[b]	12,185 + 426 = 12,611[c]	73.94%	4,445	26.06%	17,056	100%
Corpus del español – NOW	Six billion words[d]	12,869 + 403 = 13,272	88.57 %	1,712	11.43 %	14,984[e]	100%
CORPES XXI	350 millones de formas ortográficas[f]	2,386	48.96%	2,486	51.02%	4,873[g]	100%
Spanish Web 2018 (esTenTen18)	16.9 billion words[h]	87,170	64.34%	48,304	35.66%	135,474	100%

[a]*Clitics are not counted separately: "infinitives + clitic (e.g. decirlo), -ndo forms + clitic (e.g. diciéndolo), and the word del are all one word each" (https://www.corpusdelespanol.org/hist-gen/, accessed October 20, 2021).*

[b]*Clitics are counted separately: "infinitives + clitic (e.g. decir lo), -ndo forms + clitic (e.g. diciéndo lo), and the word del (de el) are all two words each" (https://www.corpusdelespanol.org/web-dial/, accessed October 20, 2021).*

[c]*The numbers refer to 12,185 occurrences of idiota as a noun, plus 426 cases in which idiota was categorized as a "noun.PROP", a proper noun.*

[d]*Clitics are counted seperately: "infinitives + clitic (e.g. decir lo), -ndo forms + clitic (e.g. diciéndo lo), and the word del (de el) are all two words each" (https://www.corpusdelespanol.org/now/, accessed October 20, 2021).*

[e]*The actual number returned by the search for idiota in the NOW-Corpus without specification of PoS amounts to 14,985, a single occurrence being categorized as a pronoun. However, no corresponding context is returned, the page where the corresponding reference should appear remains blank. Therefore, for the purposes at hand, this single instance was removed from the total amount, since there is no way of examining it any closer in order to find out how idiota might be classified as a pronoun.*

[f]*Number given for the version 0.94 (https://www.rae.es/banco-de-datos/corpes-xxi, accessed October 20, 2021).*

[g]*One instance of idiota in the CORPES XXI is depicted as an extranjerismo, a category established in the CORPES alongside the known categories for word class, such as adjetivo, sustantivo, verbo, etc.*

[h]*Source of information: https://www.sketchengine.eu/estenten-spanish-corpus/ (accessed October 20, 2021).*

for *idiota* in the *NOW*-corpus (11.43%). Apart from being the earliest of the investigation periods, and therefore probably most strongly influenced by English, T1 is the least reliable with regard to information about word-class, since about half of all items contained in it could not be assigned to any word-class at all. The results for T2, i.e., use as an adjective in 33.59%, resemble the numbers from the *Genre/Historical*-corpus and the *Spanish Web 2018 (esTenTen18)*. T3 comes closest to the result for *idiota* in the *CORPES XXI*. This latter finding is certainly influenced by the fact that, in the analysis of the *covidiota*-data, reference was made to the definitions and classifications published by the RAE (in the *NGLE*). The *CORPES XXI* being one of the corpora provided by the RAE, it can be assumed that Pos-tagging is based on the RAE's criteria as well, while the other corpora were created to match different criteria. Also, although the numbers for T2 and T3 with regard to the adjective come close to the numbers of some of the corpora, the numbers for the usage as a noun do less so. In general, comparability has to be considered somewhat limited by the fact that the *covidiota*-data contains not only the noun- and adjective-group, but also the third category of indeterminable cases which remain without classification.

To sum up, as exemplified above, the part of speech analysis for *covidiota* is, in some respects, controversial and, in some cases, it is not at all possible to assign a token to one of the two word-classes (see section Parts of Speech: *covidiota* as an Adjective and/or a Noun). Nonetheless, an increase over time in the use of *covidiota* as an adjective can be attested (even though the exact proportion might not be entirely uncontested, given the challenges of the analysis mentioned above). While the

etymon was coined as a noun in English, it is increasingly used as an adjective in Spanish, which therefore can be considered one aspect of its integration into the target language.

From Defining to Referencing—Examples and Criteria

As **Table 7** shows, more than half of the tweets in T1 contain a definition of the term *covidiot* or *covidiota*. The most commonly found definition in English is shown in **Figure 1**.

Since this definition is usually attached to the tweet as an image file, the tweet itself offers space for either an accompanying comment, often of a metalinguistic kind, as in (29), or a comment in which the author uses the term for referencing, as in (30) or (31), either to express a statement, demand, etc. of their own, or merely to offer a sample sentence for how to use the term in Spanish.

(29) *Me gustó! Covidiota, tb funciona en castellano*
　　　3x clapping hands sign
　　　3x green heart [35]
　　　'I liked it! *Covidiota*, also works in Spanish'
　　　3x clapping hands sign
　　　3x green heart

(30) *Covidiota. De estos hay muchos en mi barrio.* [84]
　　　'*Covidiota*. Of those, there are a lot in my neighborhood.'

(31) *No seas COVIDiota.*
　　　#QuedateencasaPelotudo [136]
　　　Don't be *COVIDiota.*
　　　#QuedateencasaPelotudo ('stay at home, jerk')

In other instances, the tweet text only comprises the word *covidiota*, and is then followed by the English definition depicted above, thus offering the Spanish version of the term together with its proposed meaning and sample sentences in English. There are, however, also cases in which the tweet itself presents a Spanish definition of the term *covidiota*, which can either be a very close translation, as in (32) (for the first part of the definition), or, otherwise, show considerable deviations from the original, e.g., by adding new aspects and possible meanings, as in (33).

(32) *Covidiota:*
1. Persona estúpida que ignora el protocolo de "distancia social," ayudando a esparcir el COVID-19.
"De verdad vas a visitar a la abuela? No seas un covidiota" [36]
'*Covidiota:*
1. Stupid person who ignores the protocol of "social distance," helping to spread COVID-19. "Are you really going to visit grandma? Don't be a *covidiota*"'

(33) *Covidiot, covidiota. Aún se pueden agregar definiciones, como:*
3. aquel gobernante estúpido que no tome las medidas restrictivas necesarias para evitar la propagación del virus. [40]
'*Covidiot, covidiota.* One can still add definitions, such as: 3. that stupid governing politician who doesn't impose the necessary restrictive measures to avoid the spreading of the virus.'

In total, 42 out of the 1,491 tweets (42 out of the 136 in T1, to be more precise) analyzed for this study were accompanied by this particular English definition of the original term Engl. *covidiot*. Another two comprised either a slight variant of it, which is, however, very close to the first one in terms of content, or only include the first part of the first definition, leaving out the second half.

Spanish versions of the definition, also circulating as image files, arise soon after the English one, and are also passed on through the social media networks. They are, however, less uniform and less frequent. There are only seven tweets within the corpus which contain such a file. The most frequent one is shown in **Figure 2**. It appears only three times throughout the corpus, but has an almost identical variant (which appears twice). The remaining two have to be considered independent versions. There is only one definition among them which depicts the Spanish word *covidiota* as either a noun or an adjective, while all the others classify it as a noun.

In most cases however, Spanish definitions appear within the actual text of the tweet, i.e., typed in by the individual Twitter user. As the Spanish definitions of the term during the first days of its existence constitute individually composed texts, as opposed to a "standard" definition which is spread in parallel for the English term, they exhibit a certain range of variation. Most of them agree on one basic meaning of the term *covidiota*, which corresponds to the first half of the definitions shown above, i.e., someone who ignores the appeal to keep physical distance from others, avoid social gatherings, etc., and thus is considered to further enable the spreading of COVID-19 [see for example (34), but also (35)]. Some of these definitions also take into account

the second part of the definition, which describes a person who resorts to panic purchases and hoards large amounts of certain goods at home [e.g., example (36)]. A peculiarity of example (36) is the classification of the first sense of the word as an adjective, whereas the second is registered as a noun.

(34) *COVIDIOTA: Dícese de aquella persona idiota que no atiende las norma de higiene, cuarenta y comunidad ante la pandemia de Covid-19.* [82]
'*COVIDIOTA:* It is said of the idiotic person who does not comply with the norms of hygiene, quarantine and communality in the face of the Covid-19 pandemic.'

(35) *#COVIDIOTA: /unisex/ entiéndase por persona QUE NO SE QUEDA EN SU CASA!!! #COVID19 #AlertaCOVID19SV* [119]
'*#COVIDIOTA: /unisex/* it shall be understood as a person WHO DOES NOT STAY AT HOME!!! #COVID19 #AlertaCOVID19SV'

(36) *Covidiota: 1(adj.); dicese de la persona que debiendo quedarse en su casa sale a boludear y nos caga a todos.*
2 (sust.) Persona que acapara bienes, especialmente papel higienico como si el COVID19 diera diarrea eterna.
Que estas esperando RAE?????
@RAEinforma [44]
'*Covidiota: 1(adj.);* it is said of the person who, although he/she should stay at home, goes out to piss about and messes things up for all of us. 2 (noun) Person who hoards goods, toilet paper in particular, as though COVID19 caused eternal diarrhea. What are you waiting for, RAE?????'
@RAEinforma

As has already been pointed out, in a number of uses, more than one of the categories indicated in **Table 7** are present in the tweet, as in (37). Here, the speaker first makes a comment *about* the term *covidiota* (regarding its origin and form of appearance), which is a metalinguistic comment, and then gives a short definition of the term. Thus, this tweet qualifies for "containing a metalinguistic comment" as well as for "containing a definition."

(37) *En USA están usando el HT #Covidiot para los que no respetan la cuarentena. Ya está su equivalente en español #Covidiota* [85]
'In the USA they're using the HT *#Covidiot* for those who don't respect quarantine. Now there's its equivalent in Spanish *#Covidiota*'

(38) *@RAEinforma #dudaRAE existe Covidiota? Alguien que ignora las advertencias en salud pública o seguridad. Persona que en tiempos de crisis, acumula innecesariamente bienes de primera necesidad sin tener en cuenta el bienestar de los demás.* [19]
'*@RAEinforma #dudaRAE* does *Covidiota* exist? Someone who ignores the warnings in public health or security. A person who, in times of crisis, unnecessarily accumulates consumer staples without considering the wellbeing of others.'

(38) is also an example in which metalinguistic use is combined with a definition. This example is particularly

noteworthy because the author mentions the Spanish language academy (RAE) and calls upon their lexicographic authority to ask for verification of the term's existence. References to the RAE can be made out in five other tweets, too [see also (36) above].

While tweets as (39) are categorized as pure definitions, there are other instances which appear to contain a kind of definition, or explanation, too. At the same time, they also represent an actual use of the term, an application of the neologism in context, with the function of referencing. Tweets (40) and (41) represent manifestations of this particular type.

(39) *Covidiota*
1. Persona estúpida que ignora el protocolo de distanciamiento social, ayudando a propagar COVID-19.
2. Persona estúpida que acapara víveres, aumentando el pánico por COVID-19 y dejando a otros sin recursos. #coronavirus #peru [107]
'Covidiota
1. Stupid person who ignores the social distancing protocol, helping to spread COVID-19.
2. Stupid person who hoards groceries, increasing the panic caused by COVID-19 and leaving others without resources. #coronavirus #peru'
(40) *bro, los covidiotas son los que salen sin tener la necesidad, a fiestas o reuniones, tu saliste por trabajo, eso no te hace covidiota* [216]
'*bro*, the *covidiotas* are those who go out without necessity, to parties or gatherings, you went out to work, this doesn't make you *covidiota*'
(41) *busco covidiota q me invite a una peda d Halloween*
Zany face [883]
'I'm looking for a *covidiota* who invites me to a Halloween booze-up'
Zany face

Both (40) and (41) are individual Spanish texts which contain the term *covidiota* and thus qualify for the category "referencing." On the other hand (40) very explicitly defines the author's interpretation of the term, whereas (41) contains a more implicit clarifying element: the author implies that someone who organizes a Halloween-booze-up is a *covidiota*. This type of tweet, thus, differs from the actual ("pure") definitions in that defining the term is not all, or not even the primordial function of the utterance, but only constitutes a part of it. As it seems, in (40) the author also intends to reassure the addressee, saying that he/she is not a *covidiota*. In (41), the Twitter user expresses a wish to be invited to a Halloween party. In both examples, the definition of *covidiota* is only part of the content conveyed by the texts. Tweets of this kind are therefore grouped together to form the type "referencing and negotiating semantic content." Coming back to the types of alterity markers defined by Pflanz, this function can probably be said to correspond to the fourth type: explaining the neologism (*X expliqué*, Pflanz, 2014, 173–175).

Examples (42) and (43) illustrate the difference between this type and the one depicted as "referencing only." Here, the author (tacitly) assumes a certain meaning of the term *covidiota* and uses the word accordingly, without further defining, explaining or discussing it.

(42) *Ya we ya vimos que andas de covidiota* [1109]
'Alright dude, we've already seen that you're out and about as a *covidiota*'
(43) *Que oso que te de orgullo ser un/una covidiota* [1461]
'What a bummer that being a *covidiota* makes you proud'

The distinction between both categories is made because it provides a certain amount of insight regarding the proportion of tweets which actively and explicitly contribute to the collective negotiation of the meaning of this particular neologism among the Spanish-speaking Twitter community. It is, however, not always an easy distinction to make because it is not always clear where to draw the line between both types. The crucial question is whether a meaning-defining notion can be detected in the tweet or not. If a tweet constitutes a reply to another tweet, for example, it is hard to say whether a statement such as (44) should be assigned to the first or the latter category since we do not know what is meant by *eso* 'that.' But it clearly refers to an action or event known to the addressee, and it does clearly represent a positioning of the tweet's author concerning the meaning of the term *covidiota*. Consequently, it has been assigned to the type "referencing and negotiating semantic content," in line with statements as in (40).

(44) *Eso fue muy covidiota de tu parte.* [578]
'That was very *covidiota* of you.'
(45) *Sería muy cool andar de covidiota en* CITY [995]
'It would be very cool to be out and about as a *covidiota* in CITY'

In (45), it is not absolutely clear whether *being in CITY* is what would qualify as *covidiota*, or whether this would only apply to *being in CITY and behaving in a certain way*. For the latter, (45) could not be said to contain a defining element. Here, the latter option has been chosen.

CONCLUSION

The relatively small size of the corpus (1,491 tweets in total), and the selective choice of observation periods (three isolated weeks) do not allow for far-reaching conclusions or for generalizations on a larger scale. Nevertheless, the study offers insights into appropriation processes of the English term *covidiot* by the Spanish speaking Twitter community, and into developments of the resulting anglicism *covidiota* within the receiving language.

The introduction of the English-induced neologism *covidiota* into Spanish was examined with regard to several different criteria, which were grouped together into signs of newness, on the one, and indicators of integration on the other hand. Graphical highlighting, in the function of an alterity marker, descends significantly from T1 to T2, but seems to remain stable afterwards. The use of metalinguistic comments and the proportion of tweets containing only a definition of the term also

decrease considerably, but the function depicted as "referencing and negotiating semantic content" shows a growing tendency over time. With reference to Pflanz's typology of alterity markers, the results seem to imply that the frequency and variety of alterity markers decreases over the three observation periods, but some markers, such as a certain percentage of graphical highlighting and the function of explaining, remain, or even gain more significance over time.

At the same time, several aspects seem to illustrate the integration into the language use of Spanish speaking Twitter users, and, thus, a growing resemblance to an unmarked item. *Covidiota* is increasingly used as an adjective. The cases of isolated use, which is the use without integration into the structure of the Spanish sentence, become less frequent. The categories "referencing only" and "referencing and negotiating semantic content" become the relevant ones, while definitions and metalinguistic use become much less important. Finally, some recurrent patterns of usage, several kinds of verb phrases and one noun phrase, can be identified.

In order to verify and consolidate the results obtained so far, a subsequent analysis should be carried out. In order to gain statistical significance, it would of course have to be based on a larger corpus, covering a longer observation period, and, preferably, enable the use of automated analyzation tools.

Apart from this, a comparative examination of developments in the English use of the term seems promising as well as comparisons with other languages, especially other Romance languages. This could allow gaining more insight into processes of lexical innovations as they occur on social media platforms and involve the influence of English on other languages in this medium.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

REFERENCES

Adelstein, A., and Freixa, J. (2021). *Antenario*. [Antenas Neológicas – Red de neología del español]. Available online at: https://antenario.wordpress.com/ (accessed October 20, 2021).

Arndt-Lappe, S., and Plag, I. (2013). The role of prosodic structure in the formation of English blends. *Eng. Lang. Linguist.* 17, 537–563. doi: 10.1017/S1360674313000154

Attardo, S. (2020). *The Linguistics of Humor: An Introduction*. Oxford: Oxford University Press.

Authier-Revuz, J. (1995). *Ces mots qui ne vont pas de soi: Boucles réflexives et non-coïncidences du dire*, vol. I et II. Paris: Larousse.

Balteiro, I. (2012). When Spanish owns English words. *Eng. Today* 109, 9–14. doi: 10.1017/S0266078411000605

Balteiro, I. (2018). Emerging hybrid Spanish-English blend structures: '*Summergete con socketines*'. *Lingua* 205, 1–14. doi: 10.1016/j.lingua.2017.12.010

Burgers, C., and van Mulken, M. (2017). "Humor markers," in *The Routledge Handbook of Language and Humor*, ed S. Attardo (New York, NY: Routledge).

Clarín (2020a). De covidivorcios y coronabebés: la vida durante un cierre de emergencia. *The New York Times International Weekly via Clarín* (2020). Available online at: https://www.clarin.com/new-york-times-international-weekly/covidivorcios-coronabebes-vida-cierre-emergencia_0_zfPI3qWsW.html (accessed October 20, 2021).

Clarín (2020b). Covidiota, zoompleaños, coronials: cuáles son las nuevas palabras que trae el coronavirus. *Clarín*. Available online at: https://www.clarin.com/cultura/covidiota-zoompleanos-coronials-nuevas-palabras-trae-coronavirus_0_f3apTlQiP.html (accessed October 20, 2021).

Colston, H. L. (2017). "Irony and sarcasm," in *The Routledge Handbook of Language and Humor*, ed S. Attardo (New York, NY: Routledge).

CORPES XXI = Real Academia Española (2021). *Corpus del Español del Siglo XXI (CORPES XXI)*. Available online at: https://webfrl.rae.es/CORPES/view/inicioExterno.view (accessed October 27, 2021).

Crystal, D. (2004). *A Glossary of Netspeak and Textspeak*. Edinburgh: Edinburgh University Press.

Crystal, D. (2006). *Language and the Internet. Second Edition.* Cambridge: Cambridge University Press.

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. New York, NY: Routledge.

Davies, M. (2002). *Corpus del Español: Historical/Genres*. Available online at: http://www.corpusdelespanol.org/hist-gen/ (accessed October 20, 2021).

Davies, M. (2004). *British National Corpus* (from Oxford University Press). Available online at: https://www.english-corpora.org/bnc/ (accessed October 20, 2021).

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. Available online at: https://www.english-corpora.org/coca/ (accessed October 21, 2021).

Davies, M. (2012). *Corpus del español: NOW (2012–2019)*. Available online at: https://www.corpusdelespanol.org/now/ (accessed October 21, 2021).

Davies, M. (2013). *Corpus of Global Web-Based English (GloWbE)*. Available online at: https://www.english-corpora.org/glowbe/ (accessed October 28, 2021).

Davies, M. (2016). *Corpus del Español: Web/Dialects*. Available online at: http://www.corpusdelespanol.org/web-dial/ (accessed October 20, 2021).

DHLE = Real Academia Española (2013). *Diccionario histórico de la lengua española (DHLE)* [online edition 2021]. Madrid: Real Academia Española. Available online at: https://www.rae.es/dhle/ (accessed October 21, 2021).

DLE = Real Academia Española (2014). *Diccionario de la lengua española (DLE)* [online edition 2021]. Madrid: Real Academia Española. Available online at: https://dle.rae.es/ (accessed September 20, 2021).

Friginal, E., Waugh, O., and Titak, A. (2018). "Linguistic variation in facebook and twitter posts," in *Studies in Corpus-Based Sociolinguistics*, ed E. Friginal (New York, NY: Routledge).

Kachru, B. B. (1994). Englishization and contact linguistics. *World Eng.* 13, 135–154. doi: 10.1111/j.1467-971X.1994.tb00303.x

Klosa-Kückelhaus, A. (2020a). Neue Wörter in der Coronakrise – Von Social Distancing und Gabenzaun. *Sprachreport* 36, 1–5. Available online at: https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/9934/file/KlosaKueckelhaus_Neue_Woerter_in_der_Coronakrise_2020.pdf (accessed October 20, 2021).

Klosa-Kückelhaus, A. (2020b). Bilder und Metaphern im Wortschatz rund um die Coronapandemie. Available online at: https://www.ids-mannheim.de/fileadmin/aktuell/Coronakrise/klosa_Bilder_und_metaphern.pdf (accessed October 20, 2021).

Klosa-Kückelhaus, A., Park, M., and Möhrs, C. (2020a). Das Corona-Wörterbuch. Bearbeitet von Lisa Wazulin, Jessica Blödorn, Stefan M. Dettlinger. *Mannheimer Morgen* (2020). Available online at: https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/9870/file/KlosaKueckelhaus_alWadi_Park_Moehrs_Das_Corona_Woerterbuch_2020.pdf (accessed October 20, 2021).

Klosa-Kückelhaus, A., Park, M., and Möhrs, C. (2020b). Neuer Wortschatz rund um die Coronapandemie. In Leibniz-Institut für Deutsche Sprache (IDS). *OWID Neologismenwörterbuch* [electronic ressource]. Mannheim: Leibniz-Institut für Deutsche Sprache. Available online at: https://www.owid.de/docs/neo/listen/corona.jsp (accessed October 20, 2021).

Ladilova, A. (2020). "Spanische Wortbildung im Kontext der Coronapandemie," in *Corona: Krise oder Wende? Wie Krisen Kulturen verunsichern und verändern*, ed M. O. Hertrampf (PhiN-Beiheft).

Lang, M. F. (1990). *Spanish Word Formation: Productive Derivational Morphology in the Modern Lexis*. New York, NY: Routledge.

Lexical Computing CZ s.r.o (2021). *esTenTen – Spanish corpus from the web*. Available online at: https://app.sketchengine.eu/ (accessed October 20, 2021).

McPherson, F., Stewart, P., and Wild, K. (2020). *The Language of Covid 19: Special OED Update* (Webinar, 10 September 2020).

Moliner, M. (2008). *Diccionario de uso del español. Edición en DVD*. Madrid: Gredos.

Möhrs, C. (2020). Ein Wortnetz entspinnt sich um "Corona". *Aktuelle Stellungnahmen zur Sprache in der Coronakrise*. Mannheim: Leibniz-Institut für Deutsche Sprache (IDS). Available online at: https://www.ids-mannheim.de/fileadmin/aktuell/Coronakrise/IDS_Sprache_Coronakrise_Moehrs_Wortnetz_Corona.pdf (accessed October 20, 2021).

NGLE = Real Academia Española (2009). *Nueva gramática de la lengua española* [online edition]. Madrid: Real Academia Española. Available online at: http://aplica.rae.es/grweb/cgi-bin/buscar.cgi (accessed October 20, 2021).

OED = Oxford University Press (2021). *Oxford English Dictionary* [online edition]. Oxford: Oxford University Press. Available online at: https://www.oed.com/ (accessed September 20, 2021).

Pflanz, M. L. (2014). Emprunt lexical: Existe-t-il une typologie de la phase néologique? *Neologica* 8, 157–183. doi: 10.15122/isbn.978-2-8124-2999-6.p.0157

Pharies, D. A. (1987). Blending in Spanish word-formation. *Roman. Jahrbuch* 38, 271–289. doi: 10.1515/9783110244946.271

Pons Rodríguez, L. (2020). 'Covidiota', 'balconazis', 'cuarenpena'... los neologismos que nos ha traído la pandemia. *Verne* in *El País*. Available online at: verne.elpais.com/verne/2020/04/07/articulo/1586246728_179666.html (accessed October 13, 2021).

Rey-Debove, J. (1978). Le métalangage. *Étude Linguistique du Discours sur le Langage*. Paris: Le Robert.

Rey-Debove, J. (1998). *La linguistique du signe. Une approche sémiotique du language*. Paris: Armand Colin.

Rivas Carmona, M., and Calero Vaquera, M. L. (2020). Pandemia y Posverdad: el impacto de la COVID-19 en la comunicación por WhatsApp. *Rev. Prisma Soc.* 31, 110–154. Available online at: https://revistaprismasocial.es/article/view/3892 (accessed October 20, 2021).

Rodríguez Abella, R. M. (2021). "Palabras para una pandemia. Algunas notas sobre las creaciones neológicas utilizadas para comunicar la enfermedad por coronavirus SARS-CoV-2," in *Contribuciones a la Lingüística y a la Comunicación Social. Tributo a Vitelio Ruiz Hernández*, eds L. Ruiz Miyares, R. M. Rodríguez Abella, A. Muñoz Alvarado, L. Chierichetti, and M. R. Álvarez Silva (Santiago de Cuba: Ediciones Centro de Lingüística Aplicada).

Rodríguez-Ponga, R. (2020). "El nacimiento de un nuevo vocabulario: consecuencias lingüísticas de la pandemia," in *Pandemia y resiliencia: Aportaciones académicas en tiempos de crisis*, 198–249, eds M. Kazmierczak, M. T. Signes, and C. Carreira Zafra (Pamplona: Ediciones Universidad de Navarra).

Roig-Marín, A. (2017). The Anglicisation of word formations: cross-linguistic blends in Spanish. *Eng. Today* 33, 50–56. doi: 10.1017/S0266078416000614

Roig-Marín, A. (2020). English-based coroneologisms. A short survey of our Covid-19-related vocabulary. *Eng Today* 2020, 1–3. doi: 10.1017/S0266078420000255

Ulašin, B. (2016). El cruce léxico en español. *AUC Philol.* 3, 163–182. doi: 10.14712/24646830.2016.42

Varela Ortega, S. (2005). *Morfología Léxica: la Formación de Palabras*. Madrid: Gredos.

Winter-Froemel, E. (2021). "Appealing though nebulous? Introducing the concepts of accessibility and usability to loanword research," *Conference talk at Language Contact in Times of Globalization 5 (LCTG5)*. University of Klagenfurt.

Constructional Borrowing From English in Hong Kong Cantonese

Brian Hok-Shing Chan*

Department of English, Faculty of Arts and Humanities, University of Macau, Taipa, China

Previous research on Cantonese-English contact in Hong Kong has focused on lexical phenomena, primarily lexical borrowing and intra-sentential, single-word code-switching (or code-mixing). Although code-switching may also involve longer English phrases, the English elements are mostly inserted into Cantonese-framed sentences in accordance with the Matrix Language Frame/MLF Model. In other words, the syntax of Cantonese appears to be largely intact despite words or phrases drawn from English. This paper underscores that in fact English syntax can be melded more intricately with lexis from both Cantonese and English, thus defying the MLF Model; however, recurrent cases are limited to three constructions so far, namely, the *which*-relative, the English PP-postmodifier, and an [NP COP P NP] sequence with an English preposition. A re-examination of these three constructions reveals that, rather than linguistic economy, they are semantically and pragmatically motivated to convey some specific meaning. Moreover, all these constructions are lexico-syntactic in the sense that they prototypically contain an English word, namely, the relativizer *which*, an English noun and an English preposition, respectively. Accordingly, these cases can also be treated as code-switching, though structural borrowing better captures the fact that some English syntactic structure is transferred. In line with Construction Grammar, these constructions are better understood as constructional borrowing in which each construction as a whole—composed of not only words from Cantonese and English but also a syntactic structure—conveys specific meaning. As for why such cases of structural or constructional borrowing are limited or partial, this paper suggests that it is more due to a soft constraint that separates English and Cantonese grammars—Hong Kong speakers still tend to convey a sense that they speak Cantonese among themselves—although they draw on linguistic resources from English. In this light, the Borrowability Hierarchy may be recast as a continuum of language separation and fluidity, which offers a more nuanced view to translanguaging.

Keywords: Cantonese-English contact, code-mixing/code-switching, constructional borrowing, lexical borrowing, structural borrowing, translanguaging

INTRODUCTION: LANGUAGE CONTACT PHENOMENA INVOLVING ENGLISH IN HONG KONG CANTONESE

Language contact between English and Cantonese has a long history. Earliest records of lexical borrowings from English are dated even before the colonization of Hong Kong by Britain in 1842, as British traders had already had businesses with Chinese entrepreneurs in Canton (see Bolton, 2003; Bauer, 2006). In the specific context of Hong Kong, language contact phenomena involving Cantonese and English have received much scholarly attention, and a lot of research has been done that documents and analyzes lexical borrowings from English and, more recently, code-switching (or code-mixing). This literature has primarily focused on how Cantonese adopts and integrates English elements at all levels of grammar, namely, phonology, morphology, and syntax. In other words, it is Cantonese which serves as the host language or the Matrix Language (Myers-Scotton, 1993, 2002) while English plays a limited role in inserting content words into the bilingual sentences or the discourse. Put in socio-cultural terms, Cantonese remains the language spoken by the majority of Hong Kong Chinese speakers, despite the depth and breadth of influence from English. In the case of lexical borrowing, most studies address the ways in which different English words are integrated into Cantonese. An early work on the topic, Chan and Kwok (1982) suggest that the process of borrowing is complete not only with an English word being phonologically adapted to Cantonese, but the English word is also assigned standard Chinese characters in writing; furthermore, it is frequently used by many speakers and eventually it becomes a word in Standard Written Chinese [e.g., 咖啡/*gaa3-fe1*[1] (coffee)]. More recent lexical borrowings from English may not have standard Chinese characters yet [e.g., *kei1-si2* (case)]; however, whether established or recent, English loanwords are mostly phonologically adapted to Cantonese [e.g., /s/ in *bus* becomes the onset of another syllable in 巴士/*baa1-si2* (bus), as Cantonese does not allow a fricative in coda position—see Bauer and Benedict, 1997 for more details], whereas English words that are not phonologically adapted have been treated as code-mixing/code-switching (Reynolds, 1985; Leung, 1987, 2001; Chan, 1992, 1998, 2021). In addition to phonotactics, multi-syllabic English words are often truncated in accordance with the typical word length of corresponding word classes in Cantonese (Luke and Lau, 2005; Li et al., 2016); that is, nouns are truncated to mono-syllabic or bi-syllabic words (e.g., *physics* becomes *fi1*; *qualification* becomes *kwo1-li2*) and verbs to mono-syllabic words (e.g., *monitor* becomes *mon1*). Apart from phonological

adaptation (including truncation), another hallmark for lexical borrowing from English is that an English word forms a compound with another Cantonese morpheme [i.e., the loan-blends such as 檸檬水/*ling4-mung1 seoi2* (lemon tea), RAP-歌/*rap go1* (rap song)—see Chan and Kwok, 1982; Wong et al., 2009]. Moreover, English loanwords involve distinctly Cantonese morphological processes [e.g., a verb is affixed by a Cantonese aspect marker, such as 肥咗/*fei4-zo2* (failed) or *check-*咗/*check-zo2* (checked)—see Wong et al., 2009]. Also implied—if not explicitly stated—is that these English loanwords appear in syntactic positions where their corresponding word classes in Cantonese appear; that is, verbs appear in predicative position, nouns appear as heads of noun phrases which are subjects or objects of a sentence, etc. Talking about morpho-syntax, these loanwords are not much different from single-word code-switching (or code-mixing),[2] except that in code-switching the English words are not phonologically adapted (i.e., they are pronounced very much like English); nor are multisyllabic English words truncated to fit into typical word lengths in Cantonese (Chan, 1992, 1998, 2021).[3] Of course, code-switching also involves longer elements or phrases from English, though they are also supposed to be inserted into a Cantonese-framed sentence (Chan, 1998; Leung, 2001). In sum, Cantonese-English code-switching largely observes the Matrix Language Frame Model (henceforth the MLF Model—Myers-Scotton, 1993, 2002, etc.) in which the Matrix Language (ML) sets the structure and word order of a code-switched sentence (*via the Morpheme Order Principle*) and provides the grammatical morphemes to that sentence (*via the System Morpheme Principle*). In light of the MLF Model, Cantonese is usually the Matrix Language (ML) and English is the Embedded Language (EL) which only inserts free and contentful morphemes—known as *content morphemes* in the MLF Model—into the code-switched sentences (Chan, 1998; Leung, 2001).

One perennial issue arising, in addition to enlarging the vocabulary and expressive power of Hong Kong Cantonese speakers, is whether English has impacted Cantonese in more extensive and profound ways, specifically in terms of morphology and syntax. To approach this issue, we start with the observation that Hong Kong bilingual speakers can and do use distinctively English morphological and syntactic features in what has been described as *code-mixing* or *intra-sentential code-switching*. For instance, in an early work, Leung (1987) noted that the English

---

Abbreviations: 1, first person pronoun; 2, second person pronoun; 3, third person pronoun; ASP, aspect marker; CL, classifier; COP, copular verb; COV, coverb; DEM, demonstrative; EMP, emphatic marker; EXIST, existential marker; FOC, focus marker; LOC, locative marker; LNK, linking particle; MOD, modal verb; NEG, negation marker; NOM, nominalizer; NUM, numeral; P, preposition/postposition; PL, plural marker; PRT, verbal particle; QUAN, quantifier; SFP, sentence-final particle.

[1]Transcriptions of Cantonese in this paper are based on *Jyut6-Ping3*—the Cantonese Romanization Scheme designed by the Linguistic Society of Hong Kong. See https://www.lshk.org/jyutping.

[2]The term "code-mixing" tends to be used in earlier works in the literature in referring to intra-sentential alternation of languages (Gibbons, 1987; Chan, 1992, etc.), but "code-switching" has emerged to become a more popular umbrella term which refers to both inter-sentential and intra-sentential alternation of languages (Chan, 1998; Leung, 2001) in alignment with influential works in the related literature (e.g., Gumperz, 1982; Myers-Scotton, 1993).

[3]There remain different views as to whether single-word items transferred from another language are code-switching or borrowing—Poplack (2018) argues that they are mostly borrowing or loanwords as long as there is morpho-syntactic integration with the host language, whereas Myers-Scotton (1993) holds that single-word code-switches are more widespread though less frequent than established loanwords. As for Cantonese-English contact in Hong Kong, the distinction between lexical borrowing and (single-word) code-switching is far from clear-cut. Since this study focuses on structural or constructional borrowing from English, I cannot delve into the issue here, but see Chan (2021) for some discussion.

plural marker quite often appears [e.g., see (1) below], even though Cantonese only has a plural marker used exclusively with pronouns [e.g., *ngo5* [first person]-*dei6* [plural] (we/us)].

(1) ngo5 duk6 zo2 gei2   go3 *chapters*
    I   read ASP several CL   chapters
    "I've read several chapters."

                                                          (Chan, 1992).

Syntactic features of English may be detected in fragments of English too [e.g., (2)].

(2) tung4   keoi5 gong2 gwaan1-jyu1 arrangement
    to/with 3   talk   about   arrangement
    for admission
    for admission
    "Talk to him/her about the arrangement(s) for admission."
                                              (Reynolds, 1985, p. 103).

In (2) above, the English noun phrase with a PP postmodifier (i.e., *for admission*) shows distinctive English syntactic structure as modifiers are largely prenominal in Cantonese NPs (Matthews and Yip, 2011, also see more discussion below). Translated into Cantonese, the English noun phrase would become one with a pre-modifier, as illustrated in (2a) below.

(2a) tung4   keoi5   gong2 gwaan1-jyu1 **jap6-hok6**
     to/with 3   talk   about   admission
     ge3   **on1-paai4**
     NOM   arrangement
     "Talk to him about the arrangement(s) for admission."
                                                  [Paraphrase of (2)].

Features of English morphology and syntax are also evident when English acts as the Matrix Language framing a code-switched sentence; in such examples, Cantonese acts as the Embedded Language inserting words or phrases into that sentence [e.g., (3)].

(3) I live with *gau2-zai2 bi4-bi1* and my *tung4-uk1*
    I live with puppy   baby   and my housemate
    "I live with baby puppy and my housemate."
                                              (Chan, 1998, p. 205).

In Cantonese, (3) would have to be expressed in a way in which *baby puppy* and *my housemate* precede—rather than follow—the verb *live*. In fact, as illustrated in (3a) below, the whole adverbial complement [i.e., *with gau2-zai2 bi4-bi1* (baby puppy) *and my tung4-uk1* (housemate)] comes before *zyu6* (live).[4]

(3a) ngo5 tung4   **gau2-zai2 bi4-bi1** tung4-maai4
     I   with   puppy   baby   and
     ngo5 **tung4-uk1** jat1-cai4 **zyu6**
     my   housemate   together   live
     "I live with baby puppy and my housemate."
                                                  [Paraphrase of (3)].

Despite features of English morphology and syntax in (1), (2), and (3), they arguably appear only when the speaker is speaking English. For (1) and (2), especially under the conception of code-switching, the speaker invoked these morphosyntactic features only *after* they had switched to English. In the Matrix Language Frame Model (Myers-Scotton, 1993, 2002, etc.), the plural marker -*s* is an *early system morpheme* activated alongside the content morpheme *chapter*, both of which are from English, the Embedded Language. In (2), the English noun phrase is an *Embedded Language Island* which is formed according to the syntax of the Embedded Language (i.e., English) rather than the Matrix Language (i.e., Cantonese), with the requirement that all words/morphemes are drawn from the Embedded Language. Notwithstanding these instances, there is little evidence in which Hong Kong speakers constantly transfer these features [i.e., plural-marking of common nouns in (1), postmodification in (2)] to Cantonese and use them with Cantonese lexical items (but see more discussion below).[5] English acting as the Matrix Language, as exemplified in (3), is possible but rare (Chan, 1998; Leung, 2001; Chen, 2005) among Hong Kong Cantonese speakers. The pattern seems more frequent and likely to be used when a Cantonese speaker is addressing a non-Cantonese speaker or answering an English question (Setter et al., 2010); or else the speakers are returnees who grew up in English-speaking countries and it is a style of their in-group talk (Chen, 2005, 2008, 2015). Although the returnees are supposed to be more marginal in their language practices and identity compared to "local Hongkongers" (Chen, 2008), their speech offers a glimpse into how Cantonese and English can be blended in more intricate ways that defy the asymmetry between the Matrix Language and the Embedded Language as conceptualized in the Matrix Language Frame Model (Myers-Scotton, 1993, 2002, etc.). This is illustrated in (4) below.

(4) kei4-sat6 *the way that you present yourself by*
    actually   the way that you present yourself by
    lei5   go3 *language*
    2   CL   language
    ji5-ging1 bei2 zo2 jat1 zung2 *arrogant* ge3 gam2-gok3
    already give ASP NUM CL arrogant NOM impression
    bei2 keoi5-dei6 le3
    give 3   PL SFP
    "Actually, the way that you present yourself by your language already gives them an impression of being arrogant."

                                              (Chen, 2008, p. 61).

It is implausible to determine which language—namely, Cantonese or English—is the Matrix Language or the Embedded Language for this code-switched sentence. On the one hand, many bound morphemes—or the so-called "system morphemes" in the Matrix Language Frame Model—are drawn from

---

[4]Notice, however, that (3a) is not exactly an SOV structure. The object noun phrase—*gau2-zai2 bi4-bi1 tung4-maai4 ngo5 tung4-uk1* (baby puppy and my housemate)—follows *tung4* (with), which is more of a verbal element than a preposition, technically known as a coverb in Sinitic languages (Matthews and Yip, 2011—see below for more details). In other words, (3a) is a serial verb construction [i.e., S V(COV) O V].

[5]Very occasionally internet users may use-*s* after a Chinese/Cantonese noun but instances like that are extremely rare and they seem to be funny and attention-seeking. For instance, a recent post on Instagram includes the word "鏡粉*s*" [literally *Mirror* (the name of a very popular boys' band in Hong Kong) + *fan* + *s*, that is, *the fans of Mirror*].

Cantonese, including the aspect marker (*zo2*), the nominalizer (*ge3*), the classifier (*zung2*), the plural marker (*dei6*), and the sentence-final particle (*le3*), suggesting that Cantonese is the Matrix Language; however, the subject noun phrase is clearly English in structure whose head noun *way* is followed by a postmodifying *that-clause* with a postverbal adverbial [i.e., *by lei-go3* (your) *language*], whereas in Cantonese modifiers are largely prenominal and adverbials are preverbal. What defies the Matrix Language Frame Model (Myers-Scotton, 1993, 2002, etc.) here is that this complex noun phrase cannot be an Embedded Language Island [refer to discussion of (2) above] since there are Cantonese determiners [*lei5-go3* (your)] within the PP adverbial. Moreover, it is important to note that the Matrix Language Frame Model is designed in such a way that English and Cantonese cannot be both the Matrix Language in this sentence—according to *the Asymmetry Principle* (Myers-Scotton, 2002, p. 9), if Cantonese is the Matrix Language, English must be the Embedded Language, and vice versa.[6]

In sum, Hong Kong bilingual speakers do not usually invoke English morphosyntax with Cantonese lexical items although they may do it occasionally. Exceptions such as (3) and (4) indicate that speakers can and do incorporate Cantonese lexical items into English syntactic structures, but very likely they are used in specific social situations which are considered to be more English-oriented. In other words, the constraint is not only a grammatical one *per se* (such as the Principles in the Matrix Language Frame Model) but also a sociolinguistic and an ideological one. That is to say, integrating Cantonese lexical items with English morphosyntax is generally deemed not appropriate when local Hong Kong Cantonese speakers converse with each other. In addition to the relative rarity of these instances— in comparison with lexical borrowing and single-word code-switching in the related literature—this is evidenced by the fact that the returnees would spontaneously shift to a more Cantonese-dominant "code-mixing style" and avoid English-dominant patterns such as (3) or (4) when talking to local Hong Kong speakers (Chen, 2008, 2015).[7] The indexical association of different "code-mixing styles"—i.e., the more Cantonese-dominant one vs. the more English-dominant one—and different groups of speakers—i.e., the local Hong Kong speakers vs. the returnees—is an ideological construction (Silverstein, 2003; Agha, 2007), and so is the distinction between the "local" and the "returnees" itself.

Putting aside the code-mixing style of returnees, the question arises as to whether the so-called local Hong Kong

speakers ever incorporate Cantonese lexical items into English morphosyntactic structures. If there is a general ban on melding Cantonese lexical items into English morphosyntactic structures, why would the speakers invoke these patterns? Gibbons (1987) observed a few individual English words that automatically bring along an English structure, which Gibbons (1987) describes as distinctive features of "the mixed code" [e.g., (5)]. An equivalent of (5) exhibits a different word order in pure Cantonese [e.g., (5a)].

(5) ji6-sap6 go3 *percent*
     Twenty CL   percent
     "Twenty percent"

                                                    (Gibbons, 1987, p. 63–64).

(5a) baak3-fan6 zi1   ji6-sap6
     Percentage NOM   twenty
     "Twenty percent"

                                                    [Paraphrase of (5)].

In a similar vein, Li (1999) argues that some English verbs introduce an English VO order [e.g., (6)], since the equivalent expression in Cantonese would be an OV structure [e.g., (6a)].

(6) *contact*   **nei5**
     Contact (V)   2 (O)
     "Contact you"

                                                    (Li, 1999, p. 15).

(6a) tung4 **nei5** **lyun4-lok3**
     with   2 (O)   contact (V)
     "Contact you"

                                                    [Paraphrase of (6)—Li, 1999, p. 16].

What is more, the VO structure is preferred over the OV structure since the former is more economical in invoking fewer syllables/morphemes and structure; for example, (6) requires less effort than (6a) in that the former invokes 2 words or 3 syllables whereas (6a) invokes 3 words or 4 syllables. However, as Li (1999) notes himself, the alternative VO order is available in Cantonese too, especially in spoken Cantonese [e.g., in relation to (6) and (6a), we can also say *lyun4-lok3 nei5* (contact you) in Cantonese].[8] More research and quantitative evidence is called for to confirm the claim that the English verbs have indeed introduced an English syntactic structure, or else these English verbs are inserted into a VO order framed by Cantonese along the lines of the Matrix Language Frame Model (i.e., Cantonese is the ML). At any rate, Li (1999) highlights the possibility of an English word bringing along some English syntactic structure, which he calls "lexicosyntactic transference" (p. 31).

The remainder of this paper reviews three patterns of code-switching which violate the general constraint on incorporating Cantonese lexical items into English syntactic structures. Instead of linguistic economy (Li, 1999), it is argued that all three constructions are motivated by some semantic/pragmatic function which is hard to fulfill in pure Cantonese in the light

---

[6] Myers-Scotton (2002) did mention the possibility of a *Composite Matrix Language* which draws from grammatical resources of the two languages participating in code-switching. Nevertheless, a Composite Matrix Language is supposed to be an outcome of *Matrix Language Turnover* in a contact situation of abrupt language shift in society or language attrition of individual bilingual speakers. These scenarios do not seem to apply to the majority of data discussed in this paper, as Cantonese has always remained the most widely spoken (home) language and most speakers are local Hong Kong bilinguals—maybe except the returnees Chen (2005, 2008, 2015) discussed.

[7] We can actually see (4) as an example of such "style-shifting" (Chen, 2008) as the speaker is trying to become more Cantonese-dominant in the middle of the sentence—the more English-dominant subject is followed by the more Cantonese-dominant predicate.

[8] It is noteworthy that Li (1999) based his discussion on data of written Cantonese in popular press, which may be more heavily influenced by Standard Written Chinese and somewhat different from spoken Cantonese.

of Cognitive Grammar (Langacker, 2008). The communicative benefit thus outweighs the ban on invoking English syntactic structures with Cantonese lexis. While all three constructions—composed of words from Cantonese and English—can be seen as instantiations of code-switching, they are better treated as a separate phenomenon since some English syntactic structure is triggered by an English word. Whereas this aspect is well-captured by the concept of lexicosyntactic transference (Li, 1999; Leung, 2010), structural borrowing seems more appropriate in highlighting the fact that a particular syntactic structure is borrowed. In the Conclusion section, I shall suggest that structural borrowing is recast as constructional borrowing according to which the structure of the code-switched or bilingual sentence itself contributes to the meaning of the code-switched or bilingual sentence, along the lines of Construction Grammar (Croft, 2001). Despite huge difference in their forms and way or degree of integration (to Cantonese), constructional borrowing and lexical borrowing or single-word code-switching are in fact the same in being the *mot juste* (i.e., the best expression—Poplack, 1988) in specific communicative contexts.

THE POSTMODIFYING *WHICH*-CLAUSE

Adopting Li's (1999) ideas, Leung (2010) discusses the postmodifying *which*-clause in Hong Kong Cantonese, suggesting that it is yet another instance of lexicosyntactic transference induced by the Principle of Economy. The construction is most likely drawn from English grammar since Cantonese relative clauses are premodifers in an NP (Yip and Matthews, 2000; Matthews and Yip, 2011). Limited as they are, the available naturalistic data (Leung, 2010, p. 71–74)—collected by the author and his friends in casual conversations and group meetings (Leung, 2010, p. 28)—show deviations from the conventional behaviors of relative clauses in English.[9] To begin with, it is actually a specific type of English relative clauses that is being borrowed. First and foremost, the relative clause mostly modifies the previous clause rather than an NP in it. In other words, *which* largely refers backwards to a clausal antecedent [14 out of 20 tokens in Leung's (2010) diary data], as illustrated in (7) below.

(7) ceoi4-fei1   [keoi5   dei6   hai6 wui2 fuk1 jan4
    unless   3   PL   COP will reply people
    ge3 gei1-kau3]_CLAUSE,
    NOM organization
    which ngo5   m4   zi1   hai6-m4-hai6
    which 1   NEG   know   COP-NEG-COP
    gam2 aa3…
    so   SFP

"Unless they are an organization who will make replies (to inquires), which I do not know is like that or not…"

(Leung, 2010, p. 71).

In (7) above, *which* introduces the speaker's comment (i.e., *which I do not know is like that or not*) on a situation rather than an entity expressed in the previous clause (i.e., *they are an organization who will make replies*). Even when the relative pronoun *which* refers to an NP in the beginning of the sentence, as in (8) below, the relative clause also always follows the matrix clause which contains the NP antecedent.[10]

(8) [zoeng1   zi2   soeng6-min6   ge3 man6-tai4]_NP
    CL   paper   on   NOM question
    go3 camp jap6-min6 ji5-ging1
    CL camp in   already
    gong2   zo2 dit1-dit1,
    discuss   ASP a-bit
    which ngo5 gok3 dak1 m4   sai2 zoi3 jiu3   recap
    which 1   think PRT NEG MOD again have-to recap
    "We already discussed the questions a bit on this sheet
    in the camp, which I think (we) do not have
    to recapitulate."

(Leung, 2010, p. 74).

There are no data in which the relative clause is embedded into the matrix clause [e.g., *The man [whom you talked to]_RELATIVE CLAUSE is her husband*]. The separation between a relativized NP and the relative clause is also possible in English, if it is supposedly a case of extraposition that happens when the relative clause is much longer or heavier than "the material that would follow it in the matrix clause if it occupied its default position following its antecedent" (Huddleston et al., 2002, p. 1066).

Leung (2010, p. 11) argues that the postmodifying *which*-clause is prompted by the Principle of Economy (Li, 1999). Assuming the *which*-relative clause becomes more integrated with the previous clause, the information content is expressed in one sentence; without the *which*-relative, the same content would have to be expressed in two completely separate sentences. Intuitively, it is true that *which* is optional in (7); that is, without *which*, the two clauses in Cantonese would sound perfectly grammatical. Moreover, Cantonese lacks a similar connective which is anaphoric to an antecedent in the matrix clause. Nonetheless, it is at best doubtful whether the *which*-relative in (7) is indeed integrated with the matrix clause and at worst misguided to treat the *which*-relative as more economical than the pure Cantonese counterpart without *which*—after all,

---

[9]Though not detailed further, in my reading of the data, the participants or speakers were most likely young university students (including the author and his friends) who organized activities for a Christian fellowship at school. In addition to the naturalistic data, Leung (2010) also conducted a grammaticality test with a list of invented samples of Cantonese-English code-switched sentences, including some items involving the *which/which-is* relative. This paper is primarily concerned with real usage of the construction and hence the findings of the grammaticality test are not considered here.

[10]Notice that in this particular example the NP is not exactly in Subject but in Topic position. An English translation of the sentence paying close attention to the word order of the example would read: [*The questions on this sheet*]_NP, (we) have *already discussed (them) a bit in the camp, which I think (we) do not have to recap.* There is another piece of data in Leung (2010) where *which* apparently refers to an NP at the beginning of a sentence, but that NP appears to be a Topic as well. In Standard English, a relative clause is not used to refer to a Topic NP, whether it is immediately following it (e.g., *Those books, which are interesting, I have read*) or separated from it (*Those books I have read, which are interesting*). This shows another deviation of *which*-relatives in Hong Kong Cantonese from the norms of Standard English.

in Li's (1999) formulation, the Principle of Economy refers rather straightforwardly to the use of fewer words/syllables. The separation of the NP antecedent and the *which*-relative in (8) further points to the rather loose connection between the *which*-relative and the matrix clause. At any rate, the Principle of Economy cannot explain why in Hong Kong Cantonese the relative clause always follows the matrix clause—that is, it never appears in the middle of it—and it modifies a clause [e.g., (7)—i.e., 14 out of 20 tokens in Leung's (2010) data] much more frequently than an NP [e.g., (8)—i.e., 6 out of 20 tokens]. What is more, only *which* has been attested in the relative clauses in Hong Kong Cantonese; in other words, relative pronouns such as *where*, *who/whom*, and the complementizer *that* are not found. The absence of these forms, however, follows naturally if the *which*-relatives in Hong Kong Cantonese are quintessentially non-restrictive, supplementary relative clauses which canonically refer to a preceding clausal antecedent and in which the only appropriate relative pronoun is *which* (Huddleston et al., 2002).

Another difference between the *which*-relative in Hong Kong Cantonese and that in Standard English lies in the optionality of *which* in the former. That is, the *which*-relative still sounds grammatical if *which* is deleted. Obviously, this would not apply for *subject relative clauses*—of which the *which*-relative clause in (7) is apparently an example[11]—in Standard English. For instance, *which* cannot be deleted in an English sequence such as *Derek wants to quit his job and become a Youtuber, which is shocking*, as it would render the second clause without a subject (i.e., **Is shocking*). The possibility of *which*-deletion in (7), however, becomes sensible, if *which* is not really moved from inside the relative clause as in generative analysis of English relatives. That is, the subject of the predicate—i.e., *hai6-m4-hai6 gam2* [roughly "(it) is like that or not)"]—can be analyzed as phonologically null as it is in Cantonese or other Sinitic languages (i.e., Chinese has been considered a *pro-drop* language in generative syntax, with *pro* being the covert pronoun—Yip and Matthews, 2000). In this light, it is likely that *which* in Hong Kong Cantonese is not an argument in the relative clause; more specifically, *which* in (7) may well be more like a connective (i.e., [*which*ᵢ *ngo5 m4 zi1* (*pro*ᵢ) *hai6-m4-hai6 gam2*] (which I do not know is like that or not)) than a relativizer binding a gap in the relative clause (i.e., [*which*ᵢ *ngo5 m4 zi1* ___ᵢ *hai6-m4-hai6 gam2*]). Similarly, in Hong Kong Cantonese, *which* sounds omissible too in an object relative clause such as (8). Parallel to (7), in (8), *which* is also tagged to the relative clause and coreferential with an empty pronoun—i.e., [*which*ᵢ *ngo5 gok3 dak1 m4 sai2 zoi3 jiu3 recap* (*pro*ᵢ)] [which I think (we) do not need to recapitulate]—rather than moved from within the relative clause—i.e., [*which*ᵢ *ngo5 gok3 dak1 m4 sai2 zoi3 jiu3 recap* ___ᵢ].

A third deviation from Standard English concerns the emergence of another variant of the construction in which the relativizer takes the form of *which-is* instead of *which* (7 out of 20 tokens in Leung, 2010). Precisely how *which-is* may have arisen from *which* remains cryptic. At any rate, in the same way as the *which*-relatives in Hong Kong Cantonese, the *which-is* variant always follows the matrix clause and is mostly anaphoric to a clausal antecedent in Leung's (2010) naturalistic data [see (9) below].

(9) keoi5 lei4 zo2 [di1   lou5-si1 zau6 hoi1-ci2
     3   come ASP QUAN   teacher   LNK   begin
     fan1-dong2-fan1-paai3]CLAUSE，
     divide party divide party
     which is zi1-cin4 mou5 ge3.
     which is before   NEG   SFP
     "After she came, the teachers had started to divide
     themselves up into different parties,
     which did not happen in the past."
                                              (Leung, 2010, p. 73).

Similar to *which*, *which-is* rarely refers to an NP antecedent, and there is only one such example in Leung's (2010) data.

(10) [ni1 bun2 je5]NP cyun4-fok1-jam1 hou2 jau5-jung6,
     DEM CL N   evangelism   EMP useful
     tung4-maai4 **which is** zi6-gei2 tai2 dou1 hou2 hou2-tai2
     and   which is self   read also EMP interesting
     "For evangelism, it is good to use this book, which is
     interesting to read on my own as well."
                                              (Leung, 2010, p. 72).

Leung (2010) suggests that *which-is* is a variant of *which* and a word in its own right. In other words, neither *which* nor *is* is an integral element in the relative (second) clause, but rather it is tagged to the relative (second) clause like a connective or conjunct (e.g., *nonetheless* or *moreover* in English—this would explain why *which-is* can be preceded by a conjunction [*tung4-maai4* (and) in (10)]. The intuition that *which-is* can be omitted in (9) and (10) lends further support to such an analysis, even though the version without *which-is* would sound less coherent (but nonetheless grammatical) in either case.

The above facts or properties of *which*-relatives point to a picture in which, precisely speaking, Hong Kong Cantonese borrows the non-restrictive, supplementary relative clause construction which modifies a clausal antecedent expressed in the matrix clause. Those *which*-relative clauses referring to an NP antecedent and the *which-is* variant are likely to be extensions from the canonical form (i.e., the sentential/clausal relative clause). Looking back at earlier literature, a similar example of sentential relative clause can be attested in Chan (1992), a possible precursor from which the *which*-relatives and their variants discussed in Leung (2010) could have evolved—notably, in this example, *which* refers backwards to a clause, and it is clearly not an argument in the relative (second) clause.

(11) [ngo5 m4   tung4-ji3 keoi5 ge3   ji3-gin3]CLAUSE，
     1   NEG   agree   3   NOM   opinion
     which does not mean ngo5 zaang1 keoi5
     which does not mean   1   hate   3

---

[11]The relative clause in (7) is a subject relative clause if we take *which* to originate from the subject position of the embedded clause [i.e., *which hai6-m4-hai6 gam2* (roughly "which is like that")] and it moves up to a higher clause [i.e., *which*ᵢ *ngo5 m4 zi1* ___ᵢ *hai6-m4-hai6 gam2* (which*ᵢ* I do not know ___ᵢ is like that or not)] as in generative grammar. See below.

"I do not agree with his/her opinions, which does not
mean I hate him/her."

(Chan, 1992).

Following this line of analysis, the current construction [CLAUSE
[*which/which is* CLAUSE]$_{CLAUSE}$] involves two separate clauses.
Consequently, it defies an account by the Matrix Language Frame
Model (Myers-Scotton, 1993, 2002) since the latter—together
with its Principles—applies to a mono-clause, which is couched
as a *CP (Complementizer Clause)* in terms of generative grammar.

If we conceptualize the *which*-relatives in Hong Kong
Cantonese as essentially non-restrictive, supplementary relative
clauses referring to a clausal antecedent, we reach a better
understanding of the meaning or motivation of the construction
beyond the Principle of Economy (Li, 1999). Crucially, the
construction enables a speaker to elaborate on a situation
presented as an objective fact (i.e., as encoded in the first/matrix
clause), and very often this elaboration represents the speaker's
personal assessment of the situation (i.e., as encoded in the
second/relative clause—see (7)–(10)). Accordingly, the English
relativizer *which/which-is* is used to introduce the speaker's
personal assessment. In this light, *which/which-is* in Hong
Kong Cantonese functions as a discourse or pragmatic marker
(Schiffrin, 1987) which gives the listener a hint of the speaker's
upcoming verbal action (i.e., giving an assessment of the first
clause). Without the discourse marker, the listener would still
be able to infer that the second clause represents the speaker's
personal assessment of the situation expressed in the matrix/first
clause, but more processing cost would be required for the
inference (Sperber and Wilson, 1995). In other words, the
which/which-is relative is not more economical in terms of the
number of words/morphemes invoked (Li, 1999). Rather, the
construction is more explicit and thus more economical in
terms of "processing cost" (Sperber and Wilson, 1995); that is,
it spares the listener's cognitive effort in inferring the speaker's
intentions and his/her ongoing message, hence functioning as a
contextualization cue (Gumperz, 1982).

AN ENGLISH NOUN AND A POSTMODIFYING PP

Based on a small corpus of naturalistic data, Chan (2015)
discusses another construction which also involves a
postmodifier, but here the postmodifier is well-integrated
with the clause, namely, a PP post-modifier in a complex NP, as
illustrated in (12) below.

(12) keoi5 dei6 hai6 [*plenary speakers* [*of* go2
     3   PL   COP   plenary speakers of DEM
     go3 *conference*]$_{PP}$]$_{NP}$
     CL conference
     "They were the plenary speakers of that conference."

(Chan, 2015, p. 18).

As mentioned above, Cantonese is head-final in NPs with
different types of premodifiers (Yip and Matthews, 2000;

Matthews and Yip, 2011).[12] In Cantonese syntax, the information
content in the postmodifying PP in (12) would be mapped into a
premodifier, as shown in (12a) below.

(12a) keoi5 dei6 hai6 [go2   go3 *conference* ge3
      3   PL   COP DEM CL conference NOM
      plenary speakers]$_{NP}$
      plenary speakers
      "They were the plenary speakers of that conference."

[Paraphrase of (12)].

Examples (12) and (12a) appear to be semantically equivalent.
In terms of the number of words or syllables the English NP
structure in (12) is not more economical than (12a), and hence
it is not likely to be prompted by the Principle of Economy
(Li, 1999). Furthermore, in a way very similar to (4) discussed
above, (12) violates the Matrix Language Frame Model (Myers-
Scotton, 1993, 2002)—whereas the system morphemes are drawn
from Cantonese, the putative Matrix Language, such as the plural
marker (*dei6*), the demonstrative (*go2*) and the classifier (*go3*),
the complex NP follows English grammar with a PP modifier,
flouting the Morpheme Order Principle. Crucially, the NP which
contains Cantonese elements (i.e., *go2 go3*) is not an Embedded
Language (EL) Island which may elude the grammar of the
Matrix Language (i.e., Cantonese) under the Matrix Language
Frame Model. In addition to word order in the complex NP,
another feature of the data clearly suggests that the NP is formed
according to English grammar; that is, an English noun may
idiomatically select a particular English preposition apart from
of which introduces the PP postmodifier [e.g., *contribution **to*** in
(13) below].

(13) keoi5 dei6 jing1-goi1 *make* [jat1-di1
     3   PL   should   make   some
     contribution to go2 go3 *academic community*]$_{NP}$
     contribution to DEM CL academic community
     "They should make some contribution to the
     academic community."

(Chan, 2015, p. 25).

Given that in most cases Cantonese-English code-switching
abides by the Matrix Language Frame model (Chan, 1998; Leung,
2001), we may wonder why a structure like (12) or (13) is invoked
in the first place.

Drawing on Cognitive Grammar (Langacker, 2008), Chan
(2015) proposes that the postmodifier structure [e.g., (12)] and
the premodifier structure [e.g., (12a)], while looking the same
in meaning, in fact convey different *construals*. That is, both
instances represent the same event, and yet the entities in the
event receive different levels of attention. In particular, in (12),
the head noun (*plenary speakers*) is preposed and foregrounded,
whereas in (12a), it is the NP modifier [*go2 go3 conference*

---

[12]Luke (1998) suggests that there are in fact postnominal relative clauses in
Cantonese, which are intuitively more frequent in spoken communication. Chan
(2015, p. 20) points out that these putative relative clauses always modify an object
NP in the matrix clause, and they may well be secondary predicates rather than
relative clauses *per se*.

(that conference)] that is foregrounded. In other words, (12) focuses on *what they did in the conference* (i.e., in answer to the question—*What did they do in that conference*?), while (12a) highlights *that conference in which they were plenary speakers* (i.e., in answer to the question—*Which conference were they plenary speakers of*?). In terms of Cognitive Grammar (Langacker, 2008), in (12), *plenary speakers* is the profile/figure which marks new or focused information whereas *go2 go3 conference* (that conference) is the base/ground which marks given or background information. On the other hand, in (12a), *go2 go3 conference* (that conference) is the profile/figure while *plenary speakers* is the base/ground. This analysis explains the fact that, in the majority of the examples in the dataset, the backgrounded NP embedded in the PP postmodifier is frequently marked by various forms of definite reference [i.e., 25 out of 33 tokens, e.g., *go2-go3* (that) in (12) and (13)]. However, the determiner of the complex NP is more variable [i.e., zero determiner in *plenary speakers* in (12) and indefinite determiner in *jat1-di1 contribution* (some contribution) in (13)].

As for the foregrounded head noun which appears before the PP postmodifier, it often cancels a possible inference of a speaker's previous utterance, as in (14) below.

(14) ngo5   gong2   [go2 go3 ***cutting*** *of* go2 gin6 *jacket*]_{NP}
     1   talk   DEM CL cutting of DEM CL jacket
     "I'm talking about the cutting of the jacket."
                                          (Chan, 2015, p. 29).

The context of example (14) was a wedding banquet in which the speaker told a friend in Cantonese that the "waist" (code-switched in English too) could be even "better," suggesting that his friend looked plump despite much work-out. But actually the speaker was referring to the *cutting* of his friend's jacket rather than his physique, hence example (14). The word *cutting*—in contrast with the listener's *body* as possibly inferred—is thus highlighted to clarify the speaker's previous comment.

Under this analysis, it makes sense to say, on a par with the *which/which-is* relatives discussed in section The Postmodifying *Which*-Clause (Leung, 2010), the postmodifying PP construction in Hong Kong Cantonese also borrows a specific type of structure or construction from English; only a certain type of postmodifier (i.e., PP) is borrowed among the full range of English postmodifiers in complex NPs (e.g., relative clauses,[13] non-finite clauses, *that*-clauses). Moreover, this PP contains an NP which marks given or background information, although, in English, the information status of this NP is more flexible and it can encode new or focused information too (e.g., *Churchill was a man of **great courage**; This is a book about **conspiracy theory**.*). Notwithstanding, variants of this pattern can be expected as language always keeps evolving. In the following example, the NP in the PP is horrendously complex, and in context it appears to encode new information. Plausibly prompted by end-weight or end focus (within the NP), crucially, this NP is marked by the

indefinite determiner [i.e., *jat1-di1* (some)] rather than the more common definite determiners.

(15) bei2 jyu4 [jat1-di1 hou5 *nice* ge3 ***packing*** *with* [**jat1-di1**
     for instance some very nice NOM packing with some
     daai6 ge3
     big   NOM
     ban2-paai4 zou6 ge3 *crossover* ge3 jat1-di1
     brand   do NOM crossover NOM some
     notebook]_{NP}]_{NP} dou1 wui5 jau5 ge3
     notebook   also MOD have SFP
     "For instance, there is some nice packing with some notebooks from crossovers
     launched by big brands (or companies)."
     (Context: In an TV interview, the interviewee, a sales manager, was talking about "crossover," which refers to enclosure of bags, notebooks or other accessories with a fashion magazine.)
                                          (Chan, 2015, p. 31–32).

An issue that remains unresolved is whether this construction of complex NP with a postmodifying PP is triggered by lexicosyntactic transference (Li, 1999). It is highly plausible that an English noun triggers the construction, as it appears in most of the examples [i.e., 27 out of 33 tokens, e.g., (12)–(15)]. This analysis could have interesting and significant implications in the language contact literature since nouns are often thought to be detached from specific syntactic patterns which involve them [for instance, English nouns are detached from a complex NP [DET N PP]], and so they are easily borrowed or transferred into different syntactic patterns of NP in another language). In this set of data, however, we see that an English noun can bring along a postmodifying PP, which is a distinctively English construction, although there are also plenty of examples in which an English noun is embedded into a typical Cantonese NP structure with premodifiers, as illustrated in (16) below.

(16) ni1   di1   hai6 [*staff* ge3   *problem*]_{NP}
     DEM   CL   COP staff NOM   problem
     "These are the problems of the staff."
                                          (Chan, 2015, p. 19).

Another complication is that there are a few examples in which the head noun is drawn from Cantonese, as illustrated in (17) below from Chan's (2015) dataset.

(17) ngo5   zou6   zo2 [siu2-siu2 **zi1-liu2 sau2-zaap6**
     1   do   ASP a-little   data collection
     on go2   di1   *tickets*]_{NP}
     on DEM QUAN   tickets
     "I have done a little data collection on those tickets."

Chan (2015) explains this pattern by resorting to *schematization*, drawing on the idea that a construction can be completely specific (e.g., an idiom in which all the words are fixed), completely schematic (e.g., a pattern of DET N in which different words may be inserted into the two syntactic slots—i.e., DET and N), or partially schematic or specific (e.g., NP *gives* NP *to* NP)—in line with Construction Grammar (Croft, 2001). A complex NP involving an English head noun and a postmodifying PP

---

[13]Recall that the *which/which-is* relatives discussed in section The Postmodifying *Which*-Clause (Leung, 2010) above are analyzed as non-restrictive and supplementary clauses following the matrix clause. They are not treated as post-nominal relative clauses in an NP in this paper.

[e.g., (12)–(15)] is seen as a partially schematic and specific construction which requires the head noun and the preposition to be drawn from English, but it has become more schematic more lately and so the head noun may be drawn from Cantonese too.

Judging from the idiomatic selection between the English counterpart of the Cantonese head noun in (17) (i.e., *data collection*) and the English preposition (i.e., *on*), there may well be another alternative account which is no less plausible. That is, even though the speaker is uttering the words in Cantonese [i.e., *zi1-liu2 sau2-zaap6* (data collection)], the English synonym (*data collection*) is also co-activated,[14] and the postmodifier PP and the particular preposition it selects (i.e., *data collection on X* but not **data collection in X*) are also triggered. Also, recall that Cantonese NPs are head-final, and a Cantonese noun presumably does not select a following PP complement; hence the PP [i.e., *on go2 di1 tickets* (on those tickets)] is most likely to be selected by an English noun though it is not chosen and outputted here. This account would allow us to maintain that the postmodifying PP—as a case of structural borrowing—is triggered lexically, even though in the minority of examples like (17) the triggering process is more indirect *via* a co-activated English noun.[15]

AN ENGLISH PREPOSITION IN A PREDICATIVE PP

The last construction discussed here which involves structural borrowing from English is a sequence of [NP COP P$_{English}$ NP] in which the preposition has to be drawn from English (Chan, 2018). In fact, the English preposition is also involved in a number of other constructions, no less the NP postmodifier discussed in the above section. However, the [NP COP P$_{English}$ NP] sequence appears most frequent in naturalistic data (e.g., the datasets in Chan, 2018); what is more, the construction is the earliest one to emerge with an English preposition in the literature. The following example is drawn from Gibbons (1987) whose data of the speech of Hong Kong University students were collected in the late 1970s.

(18) jat1   go3 *society* hai6 *within*   jat1   go3 *country* wo3
      NUM   CL society COP P   NUM CL country SFP
      "A society is within a country, isn't it?"

                                       (Gibbons, 1987, p. 61).

In Cantonese syntax, (18) would have to be expressed with a localizer (i.e., a postposition—Matthews and Yip, 2011; Chan, 2018), as illustrated in (18a) below.

---

[14]In the relevant psycholinguistic literature, it has been widely agreed that bilinguals co-activate words in both languages even though they are speaking only one language (Green, 1986; Kroll and Ma, 2017). Different patterns of language use—in either language or in different patterns of code-switching—are results of the bilinguals controlling and inhibiting resources from either languages in their output (Green and Li, 2014).

[15]I leave open other factors which may favor the triggering of a syntactic pattern *via* a co-activated word; for instance, the speakers may well be in a mode in which the non-selected language (i.e., English in this case) is more active (Green and Li, 2014).

(18a) jat1   go3 *society* (hai6)
       NUM   CL society COP/FOC
       hai2 jat1 go3 *country* (ge3) **leoi5-min6** wo3
       LOC NUM CL country NOM within/P   SFP
       "A society is within a country, isn't it?"

                                       [Paraphrase of (18)].

In addition to the localizer/postposition *leoi5-min6* (within/inside), a locative verb *hai2* has to be used, whereas the segmentally homophonous (but in a different tone) copular verb *hai6* is optional. Actually, *hai6* has become more of an optional focus marker (i.e., a sentence remains grammatical without it) except in contexts where the predicate is an NP [e.g., *Keoi5* **hai6** *ji1-saang1* (He/she is a doctor)—Matthews and Yip, 2011; Chan, 2018]. Looking at (18) again, it is not difficult to see that its structure is drawn from English grammar—it is much more similar to the structure of an English sentence [see the translation of (18)] rather than that of a Cantonese-framed sentence [i.e., (18a)]. Moreover, once again, it is difficult to assign Cantonese the status of Matrix Language here. Although the system morphemes—function words or bound morphemes such as the numeral and classifier (i.e., *jat1-go3*)—are drawn from Cantonese, the word order (i.e., NP COP P NP) of the sentence is drawn from English, thus flouting the Morpheme Order Principle (Myers-Scotton, 1993); nor is it plausible to treat English as the Matrix Language, as the function words and bound morphemes would then be supplied by English but not Cantonese.

In case the English preposition does not denote a temporal/spatial position (e.g., *for*), the sentence would be expressed by a serial verb construction in Cantonese syntax, as illustrated in (19) and (19a) below.

(19) ni1   go3 *course* hai6 *for* di1   lecturers gaa3
      DEM   CL course COP for QUAN   lecturers SFP
      "This course is actually for the lecturers."

                                       (Leung, 2001, p. 132).

      ni1   go3 course (hai6)
      DEM   CL course COP/FOC
      wai5 di1 *lecturers* *(**ji4 cit3**) gaa3
      COV CL lecturers LNK install SFP
      "This course is actually for the lecturers."

                                       [Paraphrase of (19)].

In (19a), the semantic equivalent of *for* is *wai5*, both being a marker introducing a BENEFICIARY. In Sinitic languages, however, it is often analyzed as a coverb rather than a preposition (Matthews and Yip, 2011), since it is obligatorily followed by another verb (i.e., *ji4-cit3* [to install/set up]), and, in general, they are more like verbs morphologically (e.g., most of them may take an aspect marker or a verbal particle).

Alternatively, the English preposition may be replaced by a full verb in Cantonese, as illustrated in (20) and (20a) below.

(20) jau5   jat1-bou6-fan6 jan4   dou1
      EXIST   part   person QUAN
      hai6 *from* jau5-cin2 ge3 *family*
      COP from rich   NOM family

"Some of these people are also from rich families."

(Chan, 2018, p. 51).

(20a) jau5   jat1-bou6-fan6   jan4   dou1   (hai6)
       EXIST   part   person QUAN   COP
       lei4-zi6   jau5-cin2 ge3   *family*
       come-from rich   NOM   family

"Some of these people also come from rich families."

[Paraphrase of (20)].

It is not immediately obvious whether examples (19) and (20) abide by the Matrix Language Frame Model (Myers-Scotton, 1993, 2002), since the English prepositions *for* and *from* are treated as content words (or content morphemes) on par with nouns and verbs, and they can be drawn from English, the putative Embedded Language (EL), according to the model. In a similar fashion, neither (19) nor (20) violates the word order of Cantonese, the putative Matrix Language, and accordingly the Morpheme Order Principle, if we treat the English preposition *for* or *from* as equivalent to Cantonese coverbs [e.g., *wai6* (for) in (19b)] or even verbs [e.g., *lei4-zi6* (come from) in (20b)], both of which are followed by an obligatory NP complement. Nonetheless, in view of (19b) and (20b), the English preposition could not be inserted into a Cantonese-framed sentence in (19) and (20). If the English preposition *for* were inserted along the lines of the Matrix Language Frame Model (Myers-Scotton, 1993, 2002), there should be a second predicate following the PP as in (19a). As for (20), it is doubtful as to why a preposition *from* can be inserted into a verb position [e.g., *lei4-zi6* (come from) in (20a)], with *categorial equivalence* being assumed to be a precondition for such insertion (Muysken, 2000; Chan, 2018).

There are a number of other more intricate arguments for the [NP COP P_{English} NP] construction [e.g., (18), (19), and (20)] to be considered a case of structural borrowing from English, which are not detailed here (but see Chan, 2018). For the purpose of this paper, it is sufficient to note that examples such as (18), (19), and (20) would be translated into different constructions with different syntax in Cantonese grammar [i.e., a postposition in (18a); a serial verb construction in (19a), and a verb in (20a)]. What is more, when an English preposition appears in predicative position in a Cantonese sentence, the Cantonese *hai6* appears more obligatory as a predicator, in other words, a bona fide copular verb rather than a focus/emphatic marker which is optional in Cantonese in most contexts. If *hai6* is absent, there must be another element which serves as the predicator; for instance, the English preposition itself is reanalyzed as a verb which may take a Cantonese aspect marker, or else the English preposition behaves as a *coverb* with another verb or predicate following. However, in comparison with the [NP COP P_{English} NP] sequence, these are really different constructions which are fewer in naturalistic data (e.g., the datasets in Chan, 2018) and which most likely emerged later (see detailed discussion in Chan, 2018). In a nutshell, although the copular verb is always drawn from Cantonese (i.e., *hai6*) for some unclear reason, [NP *hai6* P_{English} NP] originates from English grammar and is hence a case of structural borrowing.

As for the motivation for this construction, the Principle of Economy (Li, 1999) seems to apply—as the [NP *hai6* P_{English} NP] construction contain fewer words or morphemes than its corresponding expressions which are more in line with Cantonese grammar [e.g., (18a), (19a) and (20a)]. Putting aside other issues concerning the Principle of Economy,[16] it somehow assumes that a bilingual or code-switched construction is semantically identical to a counterpart in pure Cantonese; that is, it is only through its comparison with the Cantonese counterpart that the code-switched construction is considered more economical. However, through the lens of Cognitive Grammar (Langacker, 2008), the [NP *hai6* P_{English} NP] construction actually conveys nuanced meaning which eludes their apparent counterparts and which motivates its usage in specific communicative contexts. In other words, the pairs of sentences illustrated in (18)/(18a), (19)/(19a), and (20)/(20a) are not exact paraphrases and they in fact convey different *construals* (Langacker, 2008) of the same event (Chan, 2018). More specifically, in (18a), the English preposition *within* represents a RELATIONSHIP which links the *trajectory* [the figure or profiled/highlighted entity—i.e., *jat1-go3 society* (a society)] and the *landmark* [the ground or backgrounded entity—i.e., *jat1-go3 country* (a country)]. On the other hand, in (18a), the postposition/localizer *leoi5-min6* (within/inside) is conceptualized as a THING which is part of the *landmark* and which apparently conceptualizes a well-defined or bounded space (Chan, 2018). In (19), the PP [i.e., *for di1 lecturers* (for the lecturers)] is conceptualized as a *property* of the subject NP [i.e., *ni1 go3 course* (this course)], whereas the corresponding coverb phrase [i.e., *wai5 di lecturers* (for the lecturers)] is presented as an *adjunct* of another verb [i.e., *ji4-cit3* (to install/set up)]. Overall, (19a) encodes a more dynamic process than (19) with an additional verb or predicate (Chan, 2018). Concerning the last pair, (20a) encodes a more specific process as the verb is spelt out [i.e., *lei4-zi6* (come from)], whereas (20) sounds more schematic with the copular verb *hai6* which is vaguer in meaning. All in all, the [NP *hai6* P_{English} NP] construction is semantically or pragmatically motivated to convey a certain construal which differs from those conveyed by similar constructions in Cantonese [i.e., (18a), (19a) and (20a)].

DISCUSSION AND CONCLUSIONS

More than a century of Cantonese-English contact, concomitant with more widespread use of English and a higher English competence among Cantonese speakers (see Bolton et al., 2020 for a recent survey), has facilitated outcomes beyond the familiar cases of lexical borrowing and code-switching in Hong Kong Cantonese, in which single words or phrases may be borrowed

---

[16]The Principle of Economy (Li, 1999) has been taken for granted in this paper rather uncritically, though, obviously, there is much room for discussion. In my understanding, it is a simplistic generalization based on observations on some code-switched patterns which appear to be more economical (i.e., invoking fewer words/morphemes) than their monolingual Cantonese counterparts. However, it is not clear if code-switched constructions are always more economical than their Cantonese counterparts. Furthermore, there is little evidence for bilinguals comparing a code-switched construction and its Cantonese equivalent when they engage in code-switching.

Construction	Form (prototypical)	Meaning (prototypical)
Which/which-is relative	CLAUSE, *which/which is* CLAUSE	Introduces a personal assessment on a situation or entity expressed in the first/matrix clause [e.g., (7)–(10)].
Complex NP with a postmodifying PP	DET N$_{English}$ P$_{English}$ NP	Foregrounds the first noun and backgrounds the second noun [e.g., (12)–(14), (17)]
English preposition in predicative position	NP *hai6* P$_{English}$ NP	Profiles an atemporal RELATIONSHIP between trajectory and landmark [e.g., (18)–(20)].

or transferred from English and inserted into a Cantonese-framed sentence. That is, a grammatical structure or construction can also be borrowed from English with words or morphemes drawn from both English and Cantonese. This paper surveys three examples of such structural borrowing that have been documented in the existing literature, including the *which/which-is* relative, an NP with a postmodifying PP headed by an English preposition, and a sequence of [NP COP P NP] with an English preposition. All three constructions are deployed to convey some subtle semantic or pragmatic effect instead of encoding some new concept or information content without a proper word in Cantonese, as has often been observed to be the case for lexical borrowing or intra-sentential code-switching (Poplack, 1988; Myers-Scotton, 1993—also see Li, 2000 for Cantonese-English code-switching in Hong Kong).

Syntactically speaking, while single words or phrases from English are inserted into a Cantonese-framed sentence in lexical borrowing or intra-sentential code-switching, structural borrowing defies this generalization, as some kind of English structure is transferred. This difference is well-illustrated by the fact that the former abides by the Matrix Language Frame Model (Myers-Scotton, 1993) with Cantonese being the Matrix Language and English being the Embedded Language, but the latter constructions defy it in one way or the other. One mechanism underlying these three constructions of structural borrowing from English in Hong Kong Cantonese is lexicosyntactic transference (Li, 1999), in which an English word (i.e., *which/which-is*, an English noun, and an English preposition, respectively) is interwoven with an English grammatical structure. Construction Grammar (e.g., Croft, 2001) offers a more holistic framework which unifies lexicosyntactic transference (Li, 1999) and the respective semantic/pragmatic motivation of the constructions discussed above, seeing a construction as a sign which is a mapping of form and meaning. More concretely, these constructions are partially schematic and partially specific, and each of them conveys a prototypical semantic/pragmatic meaning. In light of this, the three cases of structural borrowing discussed here may be recast as *constructional borrowing* (**Table 1**).

What may be the implications of these cases of constructional borrowing on the broader picture of language contact in Hong Kong? In one perspective, as proposed in Chan (2015, 2018, 2021), Cantonese and English are in fact more deeply intertwined than the more familiar and documented cases of lexical borrowing or code-switching. This perspective tallies with previous suggestions that Cantonese-English code-switching is in fact a *mixed-code* (Gibbons, 1987; Li, 2000) which draws

resources from the two languages at all levels (i.e., phonological, lexical and syntactic). Though not identical, such a view also comes close to the hugely popular concept of *translanguaging*, according to which bi/multilingual speakers draw on language resources to make meaning and to communicate efficiently irrespective of the boundaries of "named languages" (Otheguy et al., 2015, 2019; Li, 2018). Whereas Cantonese-English bilinguals can certainly manipulate and mix features from both languages in their speech, taking into consideration the whole literature of Cantonese-English contact in Hong Kong, we do find that the three constructions reviewed in this paper are apparently rarer in comparison with lexical borrowing and code-switching in Cantonese-framed sentences.[17] What is more, these constructions most probably emerged later than lexical borrowing and code-switching, even though among them the [NP *hai6* P$_{English}$ NP] sequence was documented earlier in Gibbons (1987), and it presumably arose earlier than the *which/which-is* relative (Leung, 2010) and the NP with a postmodifying PP (Chan, 2015). Another point to note is that there are in fact plenty more syntactic differences between Cantonese and English and we may wonder why constructional (or structural) borrowing has been confined to only the three constructions so far.[18] In particular, in the case of *which/which-is* relative and the NP postmodifier, the structural borrowing is in a sense not complete. As discussed above, in the former case, only the non-restrictive and supplementary relative with *which/which-is* as relativizer is borrowed; in the latter, only PP postmodifiers but not other types of NP postmodifiers (e.g., relative clauses, *that*-clauses) are borrowed. While a much wider range of constructions have appeared with an English preposition (see Chan, 2018 for details), the [NP *hai6* P$_{English}$ NP] sequence remains dominant and there is little data in which a preposition, for instance, introduces a postverbal adjunct (e.g., *Daniel is doing his work-out **in** the gym.*).

All the above facts considered, these cases of constructional (or structural) borrowing are exceptions to the general tendency for English to be transferred to Cantonese on a lexical level

---

[17]Either based on the author's own intuition/experience or the smaller number of studies in which these constructions are documented.

[18]Yip and Matthews (2000, 2007) focused on syntactic transfer—akin to what I call structural (or constructional) borrowing here—between Cantonese and English of bilingual children in Hong Kong. They identified a wide range of constructions being borrowed from one language to another. Although the direction of transfer is primarily from Cantonese to English due to the dominance of Cantonese in the input and the environment, these works somehow testify to the possibilities of structural (or constructional) borrowing beyond the three constructions discussed here.

(recall discussion in section Introduction: Language Contact Phenomena Involving English in Hong Kong Cantonese). Taking a bird's-eye view on Cantonese-English contact, we may see that bi/multilingual speakers are always disposed to draw resources from their holistic linguistic repertoire regardless of the boundaries of "named languages" for efficient communication—as what Li (2018) has called the *Translanguaging Instinct*. Nonetheless, in early phases of language contact, when bi/multilingual speakers transfer materials from a foreign language, these materials are adapted in various ways to be nativized into the host or borrowing language. Such adaptations may be seen as these speakers demarcating the two languages despite the borrowing—in the specific context of Cantonese-English contact, phonological adaptation and truncation of an English word, as well as assigning Chinese characters to it on a graphical level—may be seen as ways in which Cantonese speakers turn it into "Cantonese" and accordingly draw a boundary between Cantonese and English. In code-switching, English words and phrases are not adapted phonologically and pronounced as they would be in "English" irrespective of the phonotactic constraints of Cantonese (Chan, 2021); in a way, the boundary between Cantonese and English has faded. In structural or constructional borrowing, as illustrated in the three constructions reviewed in this paper, the distinctiveness between the two languages blurs even further as a syntactic pattern may be drawn from English amidst others in Cantonese. If this is on the right track, the often-quoted Borrowability Hierarchy (Thomason and Kaufman, 1988) would receive an explanation along the continuum of language separation and fluidity—that is, different types of language contact phenomena, including lexical borrowing, code-switching (i.e., lexical insertions without phonological adaptation) and structural (or constructional) borrowing, often appear in fixed sequence one after another because the sequence reflects how bi/multilingual speakers draw boundaries between "named languages" in language contact at first but gradually relax them in a speech community. As for

now, the prevalence of lexical borrowing and code-switching over structural (or constructional) borrowing in Hong Kong Cantonese reflects the vitality and dominance of Cantonese in the language ecology of Hong Kong despite a prolonged period of contact with English.[19] However, as suggested by the three constructions discussed here, the separation of syntax between Cantonese and English is determined more plausibly by a soft constraint that is sociolinguistically or ideologically driven (i.e., Hong Kong speakers generally still tend to keep the syntax of Cantonese intact despite materials transferred from English) than by an absolute one imposed by the language faculty along the lines of the Matrix Language Frame Model (Myers-Scotton, 1993, 2002).

AUTHOR'S NOTE

In this paper, "Hong Kong Cantonese" is intended to be a general descriptive label for the kind of Cantonese spoken in Hong Kong; differences between Hong Kong Cantonese and other varieties of Cantonese (e.g., as spoken in Macau or Canton/Guangzhou) are not dealt with here.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

---

[19]Various strands of evidence testify to the vitality of Cantonese in Hong Kong, including the latest 2011 census and 2016 by-census results, in which Cantonese has been the most widely spoken home/usual language (as cited in Bolton et al., 2020, p. 454), and the ideology that Chinese people in Hong Kong should speak Cantonese among themselves (Chen, 2008). Also see Note 18 above.

REFERENCES

Agha, A. (2007). *Language and Social Relations*. New York, NY: Cambridge University Press.

Bauer, R. S. (2006). The stratification of English loanwords in Cantonese. *J. Chin. Linguist.* 34, 172–191.

Bauer, R. S., and Benedict, P. K. (1997). *Modern Cantonese Phonology*. Berlin: Walter de Gruyter.

Bolton, K. (2003). *Chinese English: A Sociolinguistic History*. Cambridge: Cambridge University Press.

Bolton, K., Bacon-Shone, J., and Luke, K.-K. (2020). "Hong Kong English," in *The Handbook of Asian Englishes*, eds W. Botha and A. Kirkpatrick (Chichester, West Sussex: John Wiley and Sons), 449–478.

Chan, B. H.-S. (1992). *Code-mixing in Hong Kong Cantonese-English Bilinguals: constraints and processes* (M.A. Dissertation). The Chinese University of Hong Kong, Hong Kong, China.

Chan, B. H.-S. (1998). "How does Cantonese-English code-mixing work?," in *Language in Hong Kong at Century's End*, ed M. C. Pennington (Hong Kong: Hong Kong University Press), 191–216.

Chan, B. H.-S. (2015). A diachronic-functional approach to explaining grammatical patterns in code-switching: postmodification in

Cantonese–English noun phrases. *Int. J. Bilingualism* 19, 17–39. doi: 10.1177/1367006913477921

Chan, B. H.-S. (2018). Single-word English prepositions in Hong Kong Cantonese: a cognitive-constructionist approach. *Chin. Lang. Discourse* 9, 46–74. doi: 10.1075/cld.17013.cha

Chan, B. H.-S. (2021). "Language contact phenomena in Hong Kong Cantonese: what after code-switching and lexical borrowing?," in *Crossing-Over: New Insights Into the Dialects of Guangdong*, eds C. L. Tong and J. Kuong (Macau: Hall de Cultura), 149–164.

Chan, M., and Kwok, H. (1982). *A Study of Lexical Borrowing From English in Hong Kong Cantonese*. Hong Kong: Centre of Asian Studies, University of Hong Kong.

Chen, K. H.-Y. (2005). "The social distinctiveness of two code-mixing styles in Hong Kong," in *ISB4: Proceedings of the 4th International Symposium on Bilingualism*, eds J. Cohen, K. T. McAlister, K. Rolstad, and J. MacSwan (Somerville, MA: Cascadilla Press), 527–541.

Chen, K. H.-Y. (2008). Positioning and repositioning: Linguistic practices and identity negotiation of overseas returning bilinguals in Hong Kong. *Multilingua* 27, 57–75. doi: 10.1515/MULTI.2008.004

Chen, K. H.-Y. (2015). "Styling bilinguals: analyzing structurally distinctive code-switching styles in Hong Kong," in *Code-Switching Between Structural and*

Sociolinguistic Perspectives, eds G. Stell and K. Yakpo (Berlin; Munich; Boston: Walter de Gruyter), 163–183.

Croft, W. (2001). *Radical Construction Grammar*. Oxford: Oxford University Press.

Gibbons, J. (1987). *Code-Mixing and Code-choice: A Hong Kong Case Study*. Clevedon: Multilingual Matters.

Green, D. W. (1986). Control, activation, and resource: a framework and a model for the control of speech in bilinguals. *Brain Lang.* 27, 210–223. doi: 10.1016/0093-934X(86)90016-7

Green, D. W., and Li, W. (2014). A control process model of code-switching. *Lang. Cogn. Neurosci.* 29, 499–511. doi: 10.1080/23273798.2014.882515

Gumperz, J. (1982). *Discourse Strategies*. Cambridge: Cambridge University Press.

Huddleston, R., Pullum, G., and Peterson, P. (2002). "Relative constructions and unbounded dependencies," in *The Cambridge Grammar of the English Language*, R. Huddleston and G. Pullum (Cambridge: Cambridge University Press), 1031–1096.

Kroll, J. F., and Ma, F. (2017). "The bilingual lexicon," in *The Handbook of Psycholinguistics*, eds E. M. Fernández and H. S. Cairns (Oxford: John Wiley and Sons), 294–319.

Langacker, R. W. (2008). *Cognitive Grammar: A Basic Introduction*. New York, NY: Oxford University Press.

Leung, K.-W. (2010). *Lexicosyntactic transference in Cantonese-English code-switching: the case of which-relatives* (M.A. Dissertation). The University of Hong Kong, Hong Kong, China.

Leung, T. T.-C. (2001). *An optimality-theoretic approach to Cantonese/English code-switching* (M.Phil. Dissertation). The University of Hong Kong, Hong Kong, China.

Leung, Y.-B. (1987). *Constraints on intrasentential code-mixing in Cantonese and English* (M.A. Dissertation). The University of Hong Kong, Hong Kong, China.

Li, D. C.-S. (1999). Linguistic convergence: impact of English on Hong Kong Cantonese. *Asian Englishes* 2, 5–36. doi: 10.1080/13488678.1999.10801017

Li, D. C.-S. (2000). Cantonese-English code-switching research in Hong Kong: a Y2K review. *World Englishes* 19, 305–322. doi: 10.1111/1467-971X.00181

Li, D. C.-S., Wong, C. S.-P., Leung, W.-M., and Wong, S. T.-S. (2016). Facilitation of transference: The case of monosyllabic salience in Hong Kong Cantonese. *Linguistics* 54, 1–58. doi: 10.1515/ling-2015-0037

Li, W. (2018). Translanguaging as a practical theory of language. *Appl. Linguist.* 39, 9–30. doi: 10.1093/applin/amx039

Luke, K.-K. (1998). The post-modification of noun phrases in Cantonese (粵語名詞組中的後置修飾語). *Fangyan (方言)* 1, 48–52.

Luke, K.-K., and Lau, C.-M. (2005). On loanword truncation in Cantonese. *J. East Asian Ling.* 17, 347–362. doi: 10.1007/s10831-008-9032-x

Matthews, S., and Yip, V. (2011). *Cantonese: A Comprehensive Grammar*. Second edition. New York, NY: Routledge.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Myers-Scotton, C. (1993). *Dueling Languages: Grammatical Structure in Codeswitching*. Oxford: Clarendon Press.

Myers-Scotton, C. (2002). *Contact Linguistics*. Oxford: Oxford University Press.

Otheguy, R., García, O., and Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: a perspective from linguistics. *Appl. Linguist. Rev.* 6, 281–307. doi: 10.1515/applirev-2015-0014

Otheguy, R., García, O., and Reid, W. (2019). A translanguaging view of the linguistic system of bilinguals. *Appl. Linguist. Rev.* 10, 625–651. doi: 10.1515/applirev-2018-0020

Poplack, S. (1988). "Contrasting patterns of code-switching in two communities," in: *Codeswitching: Anthropological and Sociolinguistic Perspectives*, ed M. Heller (Berlin: Mouton de Gruyter), 215–255.

Poplack, S. (2018). *Borrowing: Loanwords in the Speech Community and in the Grammar*. New York, NY: Oxford University Press.

Reynolds, S. (1985). *Code-switching in Hong Kong* (M.A. Dissertation). University of Hong Kong, Hong Kong, China.

Schiffrin, D. (1987). *Discourse Markers (Studies in Interactional Sociolinguistics)*. Cambridge: Cambridge University Press.

Setter, J., Wong, C. S.-P., and Chan, B. H.-S. (2010). *Hong Kong English*. Edinburgh: Edinburgh University Press.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Lang. Commun.* 23: 193–229. doi: 10.1016/S0271-5309(03)00013-2

Sperber, D., and Wilson, D. (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Thomason, S. G., and Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley, CA: University of California Press.

Wong, C. S.-P., Bauer, R., and Lam, Z. (2009). The Integration of English loanwords in Hong Kong Cantonese. *J. Southeast Asian Linguist. Soc.* 1, 251–265.

Yip, V., and Matthews, S. (2000). Syntactic transfer in a Cantonese-English bilingual child. *Bilingualism Lang. Cogn.* 3, 193–208. doi: 10.1017/S136672890000033X

Yip, V., and Matthews, S. (2007). *The Bilingual Child: Early Development and Language Contact*. Cambridge: Cambridge University Press.

Check for updates

Snakes, Sharks, and the Great Barrier Reef: Selected Use of Anglicisms to Represent Australia in the Australian German-Language Newspaper, *Die Woche*

Jaime W. Hunt*

Learning and Teaching, The University of Newcastle, Callaghan, NSW, Australia

Much research into language contact, specifically on anglicisms in German, investigates the appearance and use of English borrowings in the print media published in Germany, Austria, or Switzerland. While these publications are intended for a German-speaking audience in majority German-speaking countries, it remains to be explored what happens to English loans when they appear in a German-language publication produced in a majority English-speaking country. This raises questions about how the local environment, issues and events related to the country of publication are represented lexically to a local audience who are familiar with these concepts in English. Using both quantitative and qualitative methods, I analyze a selection of anglicisms related to Australia in a corpus of 25,147 types and 223,671 tokens from the Australian-published German-language newspaper *Die Woche*. The findings indicate that most of these anglicism types occur within the semantic fields of place and society. The frequency and distribution of various types, such as adapted and unadapted borrowings, loan translations and loan renditions, including flagged lexical units and codeswitching, appears to be determined mostly by authorship and intended audience of individual articles rather than the newspaper as a whole. Articles attributed to the *dpa* (*Deutsche Presse-Agentur* "German Press Agency") contain a higher incidence of loan translations, loan renditions, and flagging devices, particularly of proper nouns, than those attributed to local journalists. While this may allow international readers of the articles distributed *via* the *dpa* to better understand Australian events, institutions, social phenomena, and place, it may have a distancing or even alienating effect for local German speakers, situating them outside mainstream Australian society.

Keywords: anglicisms, Australia, English, German, borrowing, language contact, code-switching (CS)

BACKGROUND: ANGLICISMS IN GERMAN

Corpus analyses on anglicisms in the German print media have been ongoing since at least the late 1950s (e.g., Zindler, 1959; Carstensen, 1965; Yang, 1990; Donalies, 1992; Langer, 1996; Plümer, 2000; Adler, 2004; Onysko, 2007; Knospe, 2014). *Der Spiegel*, in particular, appears to have attracted the most research interest and has been considered a conduit of English lexical items into German

(Carstensen, 1965, p. 22) due to its modern, cutting-edge journalistic style influenced by American models such as *Time* magazine (Onysko, 2007, p. 99). One of the functions of anglicisms in news articles is to provide an accurate representation of the topic itself, its background, and the location of events taking place in an English-speaking country, in addition to providing a strong sense of the atmosphere, or local color, of the places and people that the journalistic report is about. Authors achieve this function by including the original English titles of, for example, institutions, or the terms denoting the cultural characteristics or political phenomena (Plümer, 2000, p. 259) that are often embedded within the culture and do not occur in the borrowing language (Onysko and Winter-Froemel, 2011, p. 1,553; Plümer, 2000, p. 259).

However, not all authors researching anglicisms include proper nouns in their analyses (e.g., Onysko, 2007; Burmasova, 2010; Pulcini et al., 2012; Zenner et al., 2012; Gottlieb et al., 2018). One such author is Yang (1990, p. 119), who excluded from his analysis of anglicisms in *Der Spiegel* proper nouns indicating people, places, and institutions, e.g., the names of presidents (*e.g., Carter, Reagan*), cities (e.g., *New York, London, Hollywood*), and political terms (e.g., *Commonwealth, Labor Party*, *State Department*). However, excluding proper nouns from anglicism studies dispenses with a rich source of stylistic information, semantic connotation, and uniqueness, as Plümer (2000, p. 259–266) argues. Therefore, she included the names of institutions such as *Air Force, College*, and *Buckingham Palace* as important indicators of local color in her study on the use of anglicisms in major newspapers and television news programs in Germany and France. Nevertheless, the inclusion of proper nouns needs to be considered according to certain guidelines. For example, names of people, such as *Carter* and *Reagan* mentioned above, currencies, and the names of private commercial enterprises may well be excluded from analysis because these may be used around the world and provide little information on English influence on any given language. However, the names of geographical locations, education institutions, as well as political and legal terms present an interesting area for exploration, particularly when such entities originate in the Anglosphere and are presented in both their original and translated forms. It warrants investigating to what extent translations of such terms do occur in a particular text, especially considering the overall context.

Thus far, research on anglicisms in general has relied on data obtained from publications appearing in majority German-speaking countries. However, German is also spoken as a minority or migrant language in many countries around the world, including Australia, where German-language radio broadcasts, podcasts, internet sites, and print media are produced for a local audience. This leads, then, to the research questions guiding this paper, namely, what happens in the German-language print media published outside Europe for German speakers in English-dominant countries such as Australia? How is Australia, its icons, culture, institutions, and way of life portrayed—through the use of unadapted anglicisms or other means such as loan translations of autochthonous terms?

The structure of this paper is as follows: I will first introduce the definition of *anglicism* and the terminology used, before

explaining the source and procedure of data extraction for analysis, including limitations. Then, I will provide quantitative results and explore specific examples in a qualitative manner and conclude by suggesting refinements for further research.

MATERIALS AND METHODS

This section contains a brief description of the source for the corpus under investigation, as well as the theoretical background and methods used to create the dataset for analysis. It concludes with a summary of the issues and limitations encountered.

The *Die Woche* Corpus

The corpus of 25,147 types and 223,671 tokens used in this study derives from 201 editions (June 2017–July 2021) of Australia's only surviving German-language newspaper, *Die Woche*. Founded in 1957, this weekly publication has a current readership of over 30,000 across the country (Die Woche Australien, 2017)[1] and is available in printed and electronic formats. Since its inception, the newspaper has provided the German-speaking public with articles on politics, economics, science, society and culture, and sport from Europe and Australia. It provides news on local events within the local German-speaking community and social clubs and is a means of connecting and informing the widely dispersed German-speaking community in Australia. It acts as a platform for businesses catering toward residents with a German, Austrian, or Swiss heritage, and advertises legal and health services where community members can use the German language with service providers.

In every 24-page edition of *Die Woche*, one page each is dedicated to news items specifically related to Australia and New Zealand, and it is these articles only that I used to source the lexical items in this study. Doing so not only increases the likelihood of the authors using English sources in their writing (Fiedler, 2012, p. 249), but it also provides the opportunity to explore how Australian culture and society is portrayed to German speakers living in the country.

The journalistic team consists of ten reporters based in Australia and in Europe and sources many articles from the *dpa* (*Deutsche Presse-Agentur*) "German Press Agency" which are available to media outlets internationally. This means that the newspaper has a duality of intended audiences. It contains not only articles written by local reporters for the local readership of German speakers participating in mainstream Australian society and who are exposed to Australian cultural concepts and Australian English, but also articles intended for a German-speaking audience abroad who do not have such familiarity or exposure to Australian society or concepts.

Creating the Dataset

To identify those anglicisms that refer to some specific aspect of Australia, its society, culture, geography, etc., I applied the broad definition of anglicisms proposed by Gottlieb (2005), p. 163 as "any individual or systemic language feature adapted or adopted

---

[1]https://www.diewoche.com.au/geschichte-der-zeitung/.

from English, or inspired or boosted by English models, used in intralingual communication in a language other than English." I interpreted this definition to also include what may be considered borderline anglicisms such as *kangaroo, koala*, and *emu*. None of these three is attested in the reference work on anglicisms in German, the *Anglizismen Wörterbuch* (Carstensen and Busse, 2001), while only *kangaroo* is classified as an anglicism in the *Dictionary of European Anglicisms* (Görlach, 2001, p. 174), and both *kangaroo* and *emu* are listed as having English origins in *Duden Online*.[2] Despite this, I included all three here because of their iconic representation of Australia and while they may have origins among Australia's Indigenous languages (*kangaroo* and *koala*) and Portuguese (*emu*), these terms have spread to other languages *through* English.

I broadly refer to the lexical items under investigation here as *Australia-related anglicisms*. These reflect either uniquely Australian concepts (including landscape features such as *Lake Mungo* or cultural events such as *Australia Day*, etc.), or those that appear within the broader Commonwealth (e.g., political concepts such as *Prime Minister*), or broader still, those lexical items that derive from the Anglosphere but still have a strong association with Australia (e.g., sports such as surfing) and mostly do not occur within the German-speaking language and cultural areas. This also extends to proper nouns that refer to, for example, places, universities, and organizations. These are of interest because they may appear in both unadapted and adapted forms, thus providing some insight into the results of language contact beyond unadapted loans.

I classified these anglicisms into the categories detailed in Gottlieb et al. (2018, p. 7), supplemented by Pulcini et al. (2012, p. 8). An adaption of these categories is included in **Table 1**. I then used the concordance program *AntConc* (Anthony, 2021) to search each word individually and ascertain its forms, i.e., to allow for various grammatical affixes showing gender, plurality, case, etc., and to produce a type and token count. The most common and easily identifiable anglicisms are the *unadapted borrowings*. These are items clearly imported from English and retain their English orthography and morphology, e.g., *Harbor Bridge*. Those items that are *adapted borrowings* are also easily identifiable as originating in the Anglosphere, but have become morphologically and/or orthographically integrated, e.g., *Kricket* "cricket." A *semantic loan* is an instance where an English meaning is applied to an existing autochthonous form, such as *feuern* "to fire" which has taken on the English sense "to dismiss someone from a job," in addition to the earlier senses including "to shoot (a weapon)." *Hybrids* are forms where both domestic and English components are combined to form compounds such as *Cricket-Spiel* "cricket game."

The final two categories, *loan translations* and *loan renditions*, are sometimes excluded from analyses of anglicisms because they are not easy to detect (Gottlieb et al., 2018, p. 7; Hunt, 2019, p. 30; Onysko, 2019, p. 188). *Loan translations* are those that are element-by-element translations of multi-element English words or phrases, leading to such examples as *Goldküste* "Gold Coast" and *Todesotter* "death adder." *Loan renditions*

TABLE 1 | Typology of anglicisms, adapted from Pulcini et al. (2012, p. 8) and Gottlieb et al. (2018, p. 7).

Type of anglicism	definition
Unadapted borrowings	Simple words (e.g., *browse*), multiword units (*bed and breakfast*), acronyms (*OMG*), terms originating in non-English speaking communities in the Anglosphere, e.g., *tomahawk* and *jungle,* and internationalisms, e.g., *hologram*, known to be coined in English
Adapted borrowings	Representing the same categories as those mentioned above, but showing signs of orthographic/morphological integration, e.g., Norwegian *blogg* < English "blog"
Semantic loans	Domestic words or assimilated borrowings taking on English sense/s, e.g., Italian *realizzare* "to become aware of" from English "realize."
Hybrids	Domestic compounds with at least one English component or one English productive affix, e.g., Norwegian *blogginnlegg* "blog post"
Loan translations	Unit-for-unit translations of English compounds, multi-word units or phraseological units
Loan rendition Pulcini et al. (2012, p. 8)	Word or multi-word unit which translates part of an English item and provides a loose equivalent for the other, e.g., Danish *genbruge* "to recycle," German *Wolkenkratzer* "skyscraper," Italian *marchio di fabbrica* "trademark"

are similar to *loan translations* except that only one element is directly translated and the other more loosely, e.g., *Feuerwehr Westaustralien* "Western Australian Fire Services" (literally "Fire Brigade Western Australia") and *Oberste Richterin* "High Court Judge (f.)," (literally "highest judge").

To identify the themes in which these Australia-related anglicisms are used, I labeled each lexical item according to its semantic field. That is, I grouped the items depending on their broad meaning or related subjects and labeled them accordingly. For example, I included *King Billy-Kiefer* "King Billy pine," *Mulgaschlange* "mulga snake/king brown snake" and *Penguin Parade* within the semantic field of *flora/fauna*, and *Backpacker-Steuer* "backpacker tax," *Liberal Party*, and *Working Holiday Visum* "working holiday visa" within the semantic field *political/legal*.

In addition to individual items, I also included code-switches in my analysis because they can also provide a good deal of insight into language contact situations and have particular pragmatic effects (Onysko, 2007, p. 89–91). However, there is debate surrounding the difference between borrowing and code-switching (see, e.g., Myers-Scotton, 1992; Clyne, 2003, p. 70–76; Knospe, 2014, p. 92–112; Poplack, 2018, Ch. 9; Gottlieb, 2020, p. 84–86; Schaefer, 2021, p. 573–574), and in synchronic analyses of language contact such as this, it is difficult to draw a clear distinction between the two (Gardner-Chloros, 2010, p. 195). However, in this paper, I take what might be considered a prototypical, narrow approach toward this issue. I restrict the concept of borrowing to that of single lexical items only, meaning that, in written discourse, prototypical code-switches are considered phrasal units or full sentences intended for plurilingual audiences (Gottlieb, 2020, p. 85; Matras, 2009,

---

[2]https://www.duden.de/.

p. 110–114; Poplack, 1993, p. 255–256; Winford, 2003, p. 107–108). Therefore, all instances of English lexical items appearing in the *Die Woche* corpus are treated as borrowings and are only treated as code-switches when it is clear from the context that an English user is the source and recipient of a message, such as in quotations, or when an entire sentence is in English.

At the same time, it is useful to recognize that code-switching and borrowing appear on a continuum (Clyne, 2003, p. 71) and many of the items under investigation appearing on this continuum are flagged. Referred to as *metacommunicative markers* by Fiedler (2012), or *alterity markers* by Winter-Froemel (2021, drawing on Pflanz, 2014), flagging devices in written discourse mark non-canonical lexical items through the use of quotation marks or italics (Saugera, 2012, p. 138–140) indicating that the items do not belong to the matrix clause and thus form "lexical islands" (Onysko, 2007, p. 296). Other flagging devices include paraphrasing, near-synonyms, translations and explanations in the surrounding text (Sharp, 2007, 234).

Excluded from this study are lexical items that are clearly of English origin but define concepts not specifically associated with Australia. An example of such a term is *Lockdown*, an attempt by governments to reduce the spread of the COVID-19 virus by temporarily requiring residents to stay at home and by moving most retail and business activity online. Furthermore, I excluded the names of private companies because generally they do not undergo morphological or orthographical adaptation and thus provide little insight into language contact processes.

Issues and Limitations of the Study

In analyzing the use of Australia-related anglicisms, various issues concerning not only the methodology but also the materials and typology need to be taken into account. Firstly, the corpus also contains articles on topics from New Zealand. However, to overcome this and to focus only on topics related to Australia (with which I am more acquainted and am able to identify more easily), I first found the lexical items under investigation in the printed editions of the corpus that focused on Australia before searching for them in the concordance software. It is here that I remind the reader that all quantitative results (i.e., type and token counts) should be considered indicative only, where the main aim is to investigate a few select items in-depth, not overall anglicism use in the newspaper.

Secondly, this paper, with its qualitative methodology, includes an interpretive analysis of the content. As such, it involves a degree of subjectivity based on my own personal, cultural, and linguistic background, current context as an L1 user of Australian English and citizen of that country. With such reliance on personal intuition, there is a risk of introducing bias, and it also means that the study is not replicable. However, it is also common in the field of anglicism research for the investigator to rely on a degree of native-speaker intuition in identifying and classifying anglicisms, particularly in instances involving the interpretation of texts and the function of lexemes (see, e.g., Burmasova, 2010, p. 50; Onysko, 2007, p. 173; Zindler, 1959, p. 2). Furthermore, the precise definition of many terms in the field are still debated among lexicographers, and thus are open to interpretation. I aim to be as objective as possible here,

while acknowledging that true objectivity in such an endeavor is impossible.

Thirdly, the corpus materials, as with many other similar studies, relied on the print media as a data source. As such, the source is written, highly edited, targets specific readers, and contains stories considered to be of interest to the public by the editorial team. In addition, the corpus is restricted further to only those articles making reference to Australia, and thus, is not comparable to other studies of anglicisms in the print media in general, nor is it truly indicative of the use of anglicisms in Australian German.

Finally, taking an approach beyond focusing on adapted and unadapted borrowings to include indirect loans such as loan translations and loan renditions may also be problematic. This is mostly because identifying such borrowings from a purely lexical point of view is difficult, and a keen sense of etymology is often required. Despite this, some indirect borrowings were identifiable and worthy of investigation as they provide insight into the variety of Australia-related anglicisms in the corpus and how Australia is portrayed in the newspaper overall.

FINDINGS AND DISCUSSION

This section contains both the numerical and qualitative findings of the analysis. Following each point is a brief discussion.

Type and Token Count

The total dataset of Australia-related anglicisms consists of 375 types and 1,653 tokens deriving from 261 English etymon types. Almost all lexical items in the dataset are nouns, except for one type and 53 tokens of *Down Under* (appearing both as an adverb as well as a noun). **Table 2** shows the frequency of the ten most frequent Australia-related anglicisms representing between 2 and 7% of anglicisms in the dataset. These include the two states of *New South Wales* and *Queensland* (with 124 and 95 tokens, respectively), the *Great Barrier Reef* (82 tokens), *Buschbrand* "bushfire" (72 tokens), *Aborigine* (64 tokens each), *Down Under* (53 tokens), *Buschfeuer* "bushfire" (43 tokens), *Tasmanien* "Tasmania" (40 tokens), *Australian Open* (36 tokens), and *Alice Springs* (28 tokens).

Categorizing the tokens into borrowing types (as shown in **Table 3**), reveals that unadapted borrowings are the most frequent, with two-thirds of tokens (67%) in the dataset belonging to this category. This is followed by loan translations at 11%, with the remaining categories of semantic borrowings, hybrids, adapted borrowings, and loan renditions each below 10%.

Unadapted Borrowings and Loan Translation of Proper Nouns

There are 27 instances where both unadapted and adapted forms of the same proper noun appear (see **Table 4**). Most of the adapted forms are either loan translations, e.g., *Hafenbrücke* for "Harbor Bridge" or *Weihnachtsinsel* for "Christmas Island," or partial loan translations such as *Commonwealth-Spiele* for "Commonwealth Games." Sometimes both forms occur within the same article.

TABLE 2 | Ten most frequent anglicisms in the corpus.

Type	Tokens	% Total anglicism tokens
New South Wales	124	7
Queensland	95	6
Great Barrier Reef	82	5
Buschbrand	72	4
Aborigine	64	4
Down Under	53	3
Buschfeuer	43	3
Tasmanien	40	2
Australian Open	36	2
Alice Springs	28	2

Each item also includes plural and genitive forms, rounded to the nearest whole number.

TABLE 3 | Total anglicism types and tokens for the dataset per type of borrowing, rounded to the nearest whole number.

Type of anglicism	Types	% Dataset types	Tokens	% Dataset tokens
Unadapted borrowings	214	57	1,100	67
Loan translations	59	16	190	11
Semantic loans	6	2	133	8
Hybrids	79	21	129	8
Adapted borrowings	13	3	89	5
Loan rendition	4	1	12	1

TABLE 4 | Unadapted and adapted forms of the same etymon co-existing in the dataset, per semantic field.

Semantic field	Unadapted form/etymon	Adapted form of same etymon
Cultural event	Australia Day	Australientag
	Miles Franklin Literary Award	Miles Franklin-Literaturpreis
Education institution	Australian National University	australische National University, Australische Nationaluniversität
	James Cook University	James Cook Universität
	Southern Cross University	Southern Cross Universität
	University of Sydney	Universität Sydney
	University of Technology Sydney	Technische Universität Sydney
Flora/fauna	Pale-headed snake	blassköpfige Schlange
	Tasmanian devils	Tasmanische Teufel
	Tassie devils	Tassie-Teufel
Geographical location	Christmas Island	Weihnachtsinsel
	Gold Coast	Goldküste
	Kangaroo Island	Känguru-Insel
	Mirra Mitta Bore	Mirra Mitta Quelle
Historical event	First Fleet	Erste Flotte
Infrastructure	Harbor Bridge	Hafenbrücke
Institution	Australian Museum	Australischen Museum
	Lowy Institute	Lowy-Institut
Landscape	Lake Mungo	Mungo-See
People	Queen Elizabeth II	Königin Elizabeth II
Political/legal	High Court	Oberstes Gericht/Oberster Gerichtshof
	National Party	Nationale Partei
	Supreme Court	Oberstes Gericht/Oberster Gerichtshof
Sport	Commonwealth Games	Commonwealth-Spiele
State/territory	Australian Capital Territory	Australian Capital Territorium
	South Australia	Südaustralien
	Western Australia	Westaustralien

There are four instances of loan renditions modeled on proper nouns. Interestingly, there was no difference made between the High Court (dealing with federal matters) and the Supreme Court (at the state level), with *Oberstes Gericht* and *Oberster Gerichtshof* being used synonymously in both instances. While both back-translate to "highest court," the use of the capital letter <o> here indicates that it is a proper noun (20 tokens) rather than a common noun (14 tokens). Similarly, the office of High Court Judge is referred to as *Oberste Richterin* "highest judge" (marked here for feminine gender). A similar lack of distinction between federal and state levels occurs with the use of the title *Premierminister(in)*. Here, the term is used to refer to both the *prime minister*, leader of the federal parliament, and *premier,* leader of each state's parliament.

There may be several reasons for including both unadapted borrowings and loan translations of the same proper nouns. Loan translations provide an accurate and factual representation of their etymons, whereas unadapted forms also add local color. Similarly, particularly if both forms appear in the same article, the variation may also be an aid to understanding, or simply a means of providing lexical variation. Nevertheless, the use of both forms of proper nouns, especially those referring to geographical locations, demonstrates that such categories form their own specific area outside mainstream anglicism studies, beyond the general lexicon (Burmasova, 2010, p. 159). As such, they warrant their own type of investigation outside the scope of this paper.

As for the loan renditions of proper nouns, making an unclear distinction between levels of the legal and parliamentary systems might not be important. This is especially so when they are presented to an international audience because the court proceedings, and results thereof, are the focal point in each article, and not necessarily the particular level of the legal system at which the trials take place. Similarly, the distinction in jurisdiction between a High Court Judge and Supreme Court Judge diminishes when the focus is placed on the crime or the outcome of the trial. Nevertheless, the same cannot be said for the blurred boundaries between Prime Minister, and Premier, where the former is the leader of the national parliament, and is active on a global stage, while the latter is relevant only at the state level.

Semantic Loan: *Busch* "Bush"

The lexeme *Busch* is the only example in the corpus that may be classified as a semantic loan, that is, where an autochthonous term takes on an English meaning. In the *Die Woche* corpus, there

are 138 tokens of the lexeme *Busch*, manifested mostly in lexical compounds such as *Buschbrand/Buschfeuer* "bushfire," *Buschland* "bushland," and *Buschwanderer* "bushwalker." Specifically, there are 72 tokens of *Buschbrand*, 41 tokens of *Buschfeuer* (appearing to be more closely modeled on the English compound), in contrast to the mere six tokens of the native equivalent of *Waldbrand* "forest fire."

In German, indeed also in British English, the noun *Busch* "bush" generally refers to a shrub; however, in Australian English (as well as other varieties of English spoken in Britain's former colonies, such as South Africa and New Zealand), a further layer of meaning has been added. In its entry for *bush*, the *Oxford English Dictionary Online* includes the definition:

> 9. a. . . . Woodland, country more or less covered with natural wood: applied to the uncleared or untilled districts in the former British Colonies which are still in a state of nature, or largely so, even though not wooded; and by extension to the *country* as opposed to the *towns* ("bush, n. 1", 2021).

And more precisely, in the *Australian Oxford English Dictionary*, this particular meaning is indicated as belonging to Australian English:

> 3. (*Aust.*) natural vegetation.
>   • a tract of land covered in this.
>
> 4. (*Aust.*) country in its natural uncultivated state.
> 5. (*Aust.*) rural as opposed to urban life; the country as opposed to the town ("bush, 1", 2004).

Thus, in Australian English, *bush* can refer not only to a shrub, but also forest, forested areas, and the countryside, and even includes small villages in rural areas. It is a culturally important, almost mythical location, connoting danger, hardship, and the pioneering spirit of the first white colonialists, while it retains a romantic sense of nostalgia that rests at the heart of many Australians, despite nearly 90% of them living in urban areas (Maude, 2018, p. 187). Bushfires, too, are an entrenched part of Australian life, particularly for those living in rural areas and the outer suburbs of Australia's cities. It is these additional layers of meaning that are evident in the German use of *Busch* in the corpus.

Reference to First Nations Peoples

The corpus contains 14 types and 64 tokens of the lexeme *Aborigine*. Thirteen types and 20 tokens appear as specifiers in hybrids, such as *Aborigine-Flagge* "Aborigine flag," *Aborigine-Kunst* "Aborigine art," and *Aborigine-Mann* "Aborigine man." In contrast, only three types and tokens of *Aboriginal* occur: once in the unadapted borrowing *Aboriginal Guide*, once in the campaign *Aboriginal Lives Matter*, and once in the hybrid *Aboriginal-Urbevölkerung* "Aboriginal native-inhabitants."

This provides perhaps the most striking piece of evidence for how a lexeme may diverge semantically in different languages. While the anglicism form in the *Die Woche* corpus appears to have retained its original neutral tone to refer to the original inhabitants of Australia, in modern Australian English, it has acquired negative connotations and is now considered taboo

and offensive. The preferred current terms include *Aboriginal and Torres Strait Islanders, First Nations peoples*, or *Aboriginal people* (Korff, 2020). While there may be diversity of opinion within the community over which of the above three terms most people identify as being, *Aboriginal* appears to be the most common, unless referring specifically to people of Torres Strait Island origin.

Semantic Fields

There are 22 semantic fields that I identified in the dataset. As shown in **Figure 1**, the most frequent (considered here to be those with a total token frequency of 50 or more) indicate a focus on place, and include names of states or territories (e.g., *New South Wales, Queensland*), landscape features (e.g., *Margaret River, Uluru*), and geographical locations (e.g., *Tamarama Beach, The Entrance*). This is followed by those anglicisms in politics and law (e.g., *Labor-Partei* "Labor Party," *High Court*), references to people (e.g., *Aussies, Crocodile Hunter*), flora/fauna (e.g., *King Billy-Kiefer* "King Billy pine," *Box Jellyfish*), and then sport (e.g., *Cricket, Tour Down Under*). A breakdown of semantic fields by borrowing type can be seen in **Table 5**, which indicates that the most common are unadapted borrowings (213 types), followed by hybrids (79 types), then loan translations (63 types) and adapted loans (14 types).

Focusing on unadapted anglicisms, as shown in **Figure 2**, there is a slight rearrangement in the order of semantic fields containing 50 or more tokens. Broadly, those lexical items belonging to the themes of *place* (e.g., states or territories, geographical location, and landscape) and *society* (e.g., people, politics, and sport) still dominate in the dataset. However, excluded from this list of semantic fields with 50 or more tokens is the field of flora/fauna. Here, there is a preference for hybridization and loan translation over unadapted borrowings. Those with simple descriptive titles, such *Tasmanische Teufel* "Tasmanian Devils" (14 tokens) and *blassköpfige Schlange* "pale-headed snake" (2 tokens) are more easily translated, negating the need for unadapted borrowing, whereas some animal names such as *kangaroo* are orthographically adapted (in this case to *Känguru*), while others, such as *koala* are left unadapted. Considering that there are 56 types of flora/fauna lexical items in the dataset and only 13 of them are unadapted reflects not only animal naming conventions, but may also suggest the intended audience for these lexical items are German speakers abroad *via* the *dpa* network, where clarity of meaning is preferred over preservation of the English label.

Code-Switching

The dataset includes four clear examples of code-switches. The first is an intersentential code-switch, acting as commentary, the second is written on a banner, the third is an indirect greeting in a quoted message on social media, and the fourth is a term of address in quoted speech. In three of the four cases, the code-switches appear to be a deliberate choice by the author for specific stylistic effects. The following shows each example of code-switching in the dataset, with the particular code-switched elements underlined here for emphasis, followed by an interpretation of its effect.
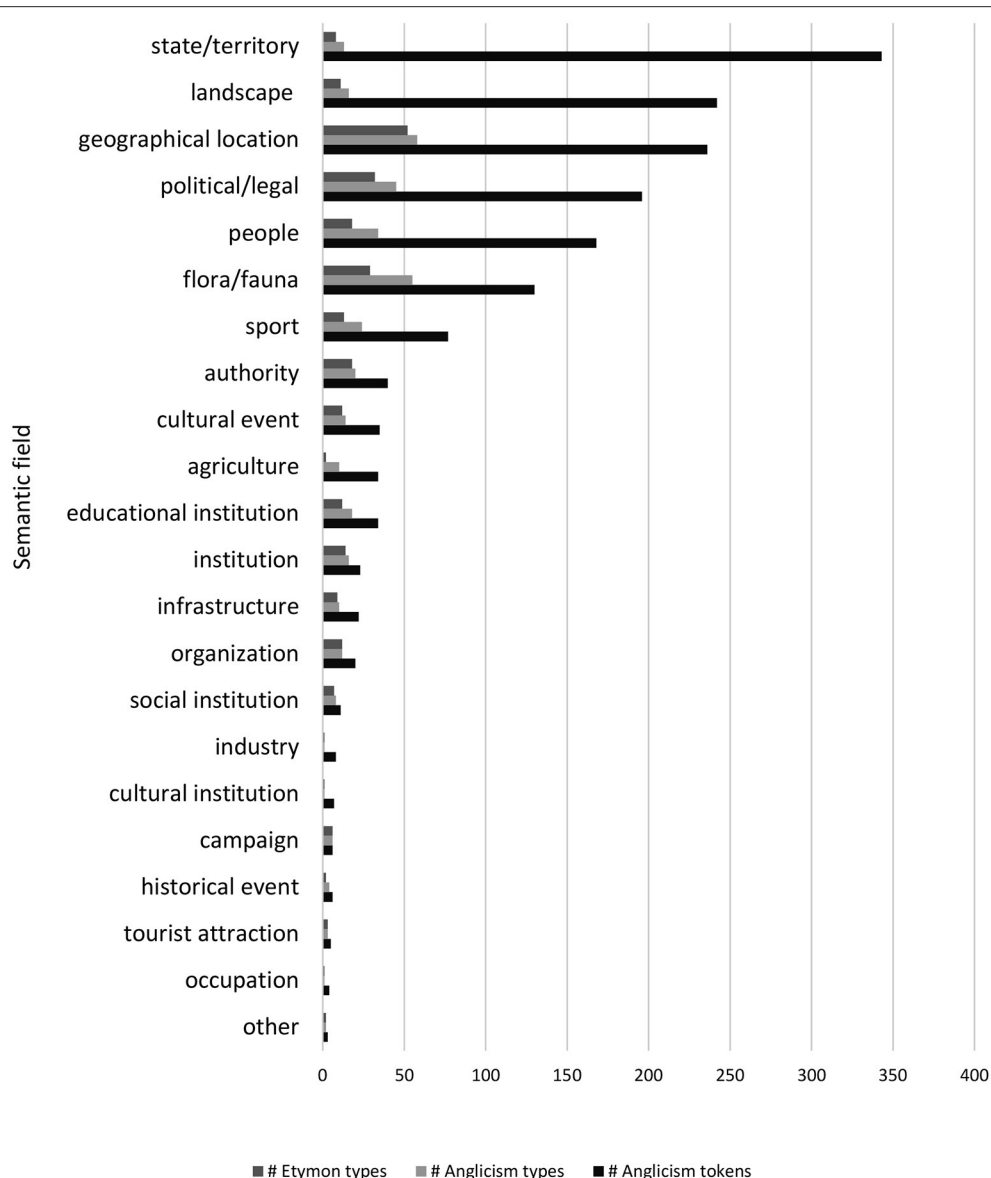
FIGURE 1 | Semantic fields by anglicism tokens and types.

The first example of a code-switch occurs in an article reporting on the 61st annual ball of the German-Austrian Society in Cabramatta (Sydney).

(1) *Und als Vorspeise mein Leibgericht: Hühnersuppe. Absolutely delicious. Zum Dessert dann Apfelstrudel und Eiscreme—wer kann diesem Menü wohl widerstehen?* (Birgit Myrach, Issue 10, 2017)

[And for entrée, my favorite dish: chicken soup. Absolutely delicious. Then for dessert, apple strudel and ice cream—who can resist this menu?]

In this article, the local reporter, Birgit Myrach, provides an account of the evening's proceedings, listing prominent community members in attendance and, as shown here, describing the evening's fare. This code-switched text functions as a direct comment by the author in her otherwise all-German prose. Furthermore, it is not flagged in any way: The English adjective phrase is not presented in quotation marks or in italics, nor is any translation offered. This suggests an assumption that the readership is familiar enough with English to understand this comment. Unlike the other articles containing code-switches below, this one stands out as being attributed to a local author reporting on a local community event for a local audience.

The second example of code-switching appeared on a sign held up by crew members aboard the cruise ship *Artania* when leaving Fremantle Harbor, south of Perth.

Semantic field	Unadapted	Hybrid	Loan translation	Adapted
Geographical location	48	4	6	
Political/legal	23	8	12	2
Authority	17	2	1	
People	12	17	5	
Institution	12		4	
Organization	12			
Cultural event	11	1	2	
Educational institution	10	3	5	
Sport	10	6		8
Flora/fauna	13	28	16	2
State/territory	8	1	3	1
Infrastructure	8		2	
Landscape	6		4	
Social institution	6	2		
Campaign	6			
Agriculture	3	7		
Tourist attraction	3			
Other	2			
Historical event	1		3	
Cultural institution	1			
Occupation	1			
Industry	1			
Total	214	79	63	13

(2) *Das Schiff verabschiedete sich mit Hornsignalen, auf den Decks winkten Crewmitglieder den Menschen im Hafen zu. ≪Thank you Fremantle≫, war auf einem Transparent mit Herz zu sehen.* (*dpa*, Issue 16, 2020)

[The ship bade farewell by blasting its horn, crewmembers waved to the people along the harbor. "Thank you Fremantle" could be seen on a banner along with a heart].

The German-operated ship had been docked in quarantine in March–April 2020 because of multiple COVID-19 cases on board, many of which required hospitalization. The city of Perth had organized for ∼850 passengers and crew members to alight the ship and, under strict quarantine conditions, to board charted flights to Germany. The remaining 400 on board, mostly crew members, departed for Germany on the ship.

Similar to example (1), the recipient of the message is important, but instead of the reader of the *newspaper* receiving the written information, readers of the sign *reported about within the article* are significant. The code-switched text in (2), "Thank you Fremantle," was intended for locals waving farewell to the cruise ship on the dock as it left port in Western Australia. Here, rather than using English to provide a commentary on the article's content, the unnamed *dpa* author gives an accurate portrayal of the sign's simple message that would not be as accurately purveyed had he/she translated the sign into German. Furthermore, most German speakers would arguably understand the phrase "thank you," thus negating the need for a translation.

The third example of code-switching occurred in a message on the social media platform Twitter by Scott Morrison, the Prime

Minister of Australia, in a tweet congratulating Boris Johnson on his re-election:

(3) ≪*Sag G'day zu den stillen Briten von uns≫, fügte er hinzu und bezog sich damit auf die Unterstützer Johnsons.* (*dpa*, Issue 50, 2019)

["Say g'day to the quiet Britons for us," he added, referring to Johnson's supporters].

Instead of the entire tweet being translated into the Standard German "*Grüß den stillen Briten von uns*" (literally "greet the quiet Britons from us") or similar, the original greeting "g'day," an abbreviated form of "good day," is retained. The inclusion of the code-switch appears to be for stylistic purposes, that is, to provide the desired connotation intended in the English original form. In using this informal greeting, the prime minister depicts himself as being down-to-earth, friendly, and relaxed (positive qualities that are strongly associated with the stereotypical Australian character), especially when compared to the more formal variant. Using this casual greeting to the head of the British parliament also portrays the stereotypical Australian value of egalitarianism and as such, provides an excellent example of local color. However, at the same time, the effect of this casual greeting may be lost on many users of German outside Australia because it requires familiarity with Australian culture and its variety of English to understand the connotations evoked.

The fourth code-switch is the only one appearing in quoted speech. It is in an article containing various critiques of Australia's policy on energy production and the environment:

(4) "*Prime Minister, sagen Sie das den schon 17 Ländern auf der ganzen Welt, die heute mehr als 90% ihres Stroms aus erneuerbaren Energien erzeugen,*" *bittet Professor Flannery.* (Unattributed, Issue 42, 2018)

["Prime Minister, tell that to the 17 countries around the world who already produce more than 90% of their power from renewable sources" requests Professor Flannery].

Here, retaining the address term *Prime Minister* within the translated quote, has the effect of providing a sense of local color by using the English title for the head of the Australian government. While translated equivalents of the title exist in German, such as *Premierminister* or *Ministerpräsident*, neither of those would represent the Australian parliamentary system as accurately as the English.

Overall, there are few instances of code-switching in the corpus. Out of those that do occur, only one, as shown in example (1) might be considered a *true* code-switch in the sense that it appears in a text that has presumably not been translated from English. While examples (2–4) are each unique in that they are from a handwritten sign directed toward local Australians, a written quote to an English speaker, and a spoken quote to an English speaker, respectively, they are all similar in that they are examples of code-switches appearing in translated texts. And as such, they appear to be more deliberate choices by the author as an indication of local Australian color, particularly (2) and (3) which are attributed to the *dpa*.

Six campaign slogans in the corpus form a particular subset of code-switches and all appear in articles written by *dpa* journalists. Therefore, it can be assumed that these authors must also
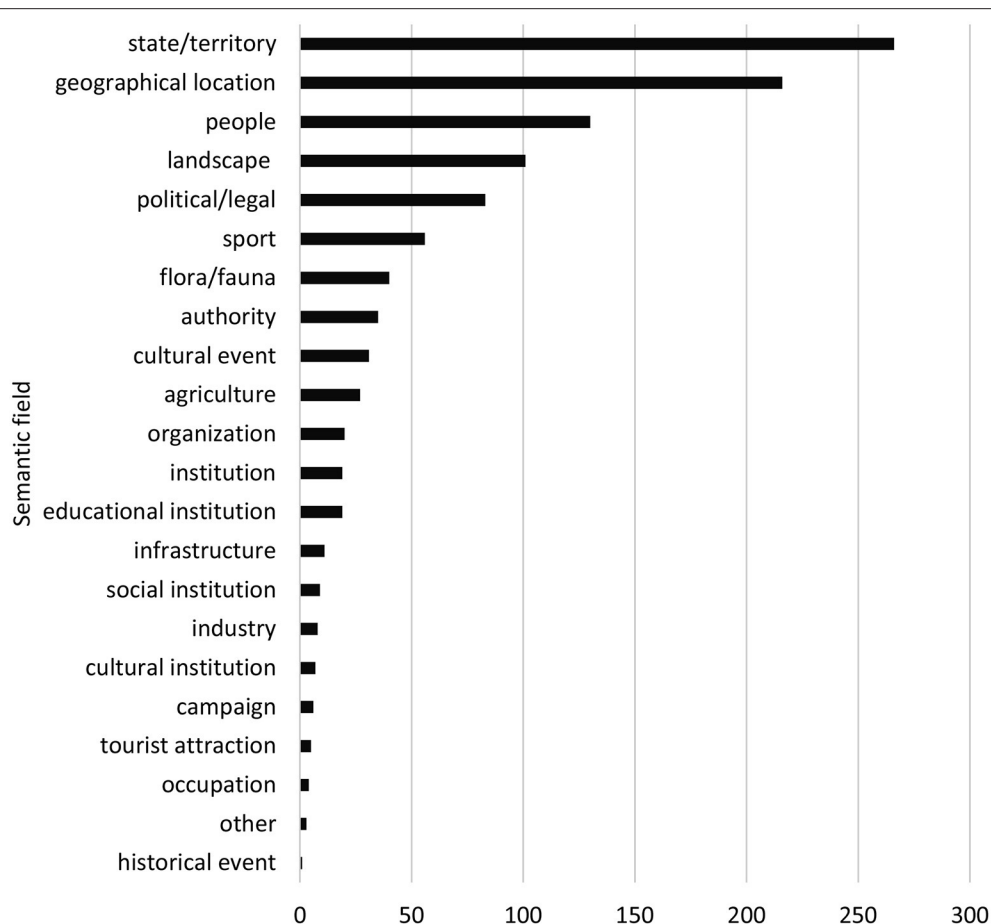
FIGURE 2 | Unadapted anglicism tokens per semantic field.

consider readers outside Australia. All examples appear once only in the corpus, apart from that shown in (5) which appears twice in the same article. They are as follows:

(5) ≪*Fly me to the Supermoon*≫ (*dpa*, Issue 21, 2021)

- this is in reference to a campaign slogan for an airline offering chartered flights from Sydney out over the Pacific Ocean to get a better view of the blood moon and lunar eclipse appearing in May 2021;

(6) ≪*Aboriginal Lives Matter*≫ (*dpa*, Issue 23, 2020)

- this is a local variant of the *Black Lives Matter* movement;

(7) #*aapneedsyou* (AAP braucht Dich) (*dpa*, Issue 38, 2020)

- a hashtag with the aim of crowdsourcing funds to save the Australian Associated Press (AAP), which was in financial difficulty and under threat of administration;

(8) ≪*Holiday Here This Year*≫ (*Urlaub hier dieses Jahr*) (*dpa*, Issue 42, 2020)

- a tourism campaign slogan encouraging Australians to holiday within their own country;

(9) ≪*pee, poo and (toilet) paper*≫. *Auf Deutsch:* ≪*Pipi, Kacke und (Toiletten-)Papier*≫ (*dpa*, Issue 14, 2020)

- an awareness campaign on what does (and more specifically, does not) belong in sewerage systems; and

(10) ≪*Save the Tasmanian Devil*≫ (≪*Rettet den Tasmanischen Teufel*≫) (*dpa*, Issue 13, 2017)

- a program designed to prevent the Tasmanian Devil from dying out due to Devil Facial Tumor Disease.

An interesting feature of these campaign slogans is that (5) and (6) are not translated, whereas (7–10) are. Example (5) is a play on the song title *Fly me to the Moon*, popularized by many popular singers including Frank Sinatra, and thus, would appear to be familiar enough to the German-speaking audience abroad as to not require translating. Similarly, example (6), a local variant of the global *Black Lives Matter* movement, would undoubtedly be familiar with a German-speaking audience abroad as well (The English form *Black Lives Matter* slogan also appears twice in Issue 24, 2020). However, the remaining slogans are much more specific to the local context and are translated. These campaign slogans, as longer strings of text flagged with quotation marks and, in some cases, translations, clearly do not belong to the

matrix language. As such, the sense of local color that they provide and as their translated forms appear to be for the benefit of those outside Australia who are reading the articles through the *dpa* network.

Finally, considering that prototypical code-switches are intended for plurilingual audiences, the untranslated code-switches as exemplified above indicate an expectation that the audience is proficient enough in English to understand them. Thus, these code-switches may be included as a stylistic device to add local color and specific nuance and meaning to the texts. However, these effects are diminished in those code-switches that are translated. Indeed, these may not be considered full code-switches as such because the apparent necessity of providing a translation assumes that the audience lacks the English proficiency to make sense of the original phrasal units. Therefore, the sense of local color, while still present, is no longer as strong because the assumption about reader proficiency appears to have changed.

Flagged Lexical Units

Multiple anglicisms appearing in the dataset are marked with quotation marks. For example, the anglicism *Aussies* appears twice flagged within the same issue, as shown in (11) and (12):

(11) *Dieses Jahr aber sind die* ≪*Aussies*≫ *mit etwas anderem beschäftigt.* (*dpa*, Issue 04, 2020)

[But this year, the Aussies are occupied with something else]

(12) *Wie geht das Leben für die* ≪*Aussies*≫ *weiter*? (*dpa*, Issue 04, 2020)
[How will life go on for the Aussies?]

However, in a separate issue, it is unflagged:

(13) "*Ich frage für 40.000 gestrandete Aussies, die nicht nach Hause in ihr Land kommen können, nachdem sie seit zwölf Monaten ohne Unterstützung oder Essen oder auch nur ein Dach über ihrem Kopf leben*" (Barbara Barkhausen, Issue 14, 2021)
[I am asking for 40,000 stranded Aussies who cannot come home to their own country after they have been living without support or food or even a roof over their heads for 12 months].

This excerpt here is from an article by local reporter Barbara Barkhausen and is not attributed to the *dpa*. Hence, it is a clear example of a local author writing for local readers and for whom *Aussie* is a common term. The remaining examples of flagged lexical units all appear in articles attributed to the *dpa*.

In some cases, complex flagging occurs. Not only orthographically marked, but also reformulated is the example of the surrounding region of Canberra, the nation's capital city:

(14) ≪*Australian Capital Territorium*≫ (ACT)—*so heißt die Region um die Hauptstadt* (*dpa*, Issue 05, 2020)
["Australian Capital Territory" (ACT)—that is the name of the region around the capital city].

Indeed, example (14) could also be viewed as being triple marked because it appears in quotation marks, it is a partial loan

translation (i.e., by translating the third element of the name from *Territory* to *Territorium*), and it is followed by an explanatory statement. That such marking occurs at all indicates a view from the outside, that is, from the point of view of a *dpa* article with an international audience in mind, because the local readership would not need such clarification.

Two forms of *Tasmanian Devils* appear in the corpus. There are *two* tokens flagged with double quotation marks and one token of the hypocoristic ≪*Tassie Devils*≫. Both the partial translation of *Tassie-Teufel* (2 tokens) and the loan translation *Tasmanische Teufel* (14 tokens) remain unflagged.

Two further examples of multiple flagging through quotation marks and as direct translation refer to two people treated as national heroes for particular acts of rebellion and defiance. They are *Egg Boy* (Issue 22, 2019) and *Trolley Man* (Issue 46, 2018). *Egg Boy* refers to Will Connolly, who broke an egg onto the head of Senator Fraser Anning who notoriously drew a connection between Muslim immigration and terror attacks. At first, only the direct translation ≪*Jungen mit dem Ei*≫ (literally "boy with the egg") appears in the headline in quotation marks. Later in the same article, the anglicism appears flagged both with a preceding explanatory text and then is followed by the direct translation. Here, there also is the indication that the audience for this article is not limited to Australia, since the locational context, indicated by the preposition phrase "*In seiner Heimat*" "in his homeland" appears targeted to readers outside Australia:

(15) *In seiner Heimat nennt man ihn* ≪*Egg Boy*≫, *den* ≪*Jungen mit dem Ei*≫. (*dpa*, Issue 22, 2019).
[In his homeland he is called "egg boy," the "boy with the egg"].

Similarly, *Trolley Man* is flagged multiple times by quotation marks as well as a loose translation. The nickname was given to a member of the public, Michael Rogers, who used a shopping trolley to prevent a knife-wielding terrorist from continuing his rampage through a busy pedestrian mall in Melbourne. In the headline, the nickname appears without any translation or explanation:

(16) *Melbournes* ≪*Trolley Man*≫—*Ein Held muss vor Gericht* (*dpa*, Issue 46, 2018).
[Melbourne's "Trolley Man" — A hero has to go to court].

Later, after several sentences explaining how the homeless man came to receive this nickname, the label appears again in quotation marks, but this time followed by a loose translation:

(17) *Rogers wurde für seinen Einsatz als* ≪*Trolley Man*≫ *(der Mann mit dem Einkaufswagen) gefeiert.* (*dpa*, Issue 46, 2018)
[Rogers was hailed for his actions as the "Trolley Man" (the man with the shopping trolley")].

In addition to orthographical flagging, some lexical items receive multiple linguistic flags. For example, *Australia Day* occurs 10 times in the corpus. In addition to when it appears unflagged as *Australientag*, the following show some of the various flagging devices given to this item:

A definition (underlined here for emphasis) in the sentence following the term:

(18)   *… sollte es demnach auch zu den Feiern zum Australia Day am Samstag brütend heiß werden. Am Nationalfeiertag gedenken die Australier…* (*dpa*, Issue 04, 2019)
[… it should therefore also be scorching hot for the celebrations for Australia Day on Saturday. On the national public holiday, the Australians commemorate …].

A definition preceding the term within the same sentence:

(19)   *Mit Partys und Protesten hat Australien am Freitag seinen Nationalfeiertag begangen, den Australia Day.* (*dpa*, Issue 04, 2018)
[On Friday, Australia celebrated its national public holiday, Australia Day, with parties and protests].

A longer explanatory text following the term, providing further situational context of the article:

(20)   *Am Sonntag war Australia Day—ein umstrittener Feiertag in Australien, weil er zum Ärger der Ureinwohner an die britischen Siedler erinnert.* (*dpa,* Issue 04, 2020)
[Sunday was Australia Day—a controversial public holiday in Australia because, much to the anger of the Indigenous people, it is a reminder of the British settlers].

With an explanation in a new sentence after the term:

(21)   *Jetzt also der Australia Day. Er erinnert daran, wie die Briten 1788 in Sydney landeten, um dort eine Kolonie für Häftlinge zu errichten.* (*dpa*, Issue 04, 2020)
[And now Australia Day. It commemorates how the British landed in Sydney in 1788 to establish a colony for convicts].

Triple-flagging with quotation marks followed by a translation, and then explanation:

(22)   *Am ≪Australia Day≫ (Australientag) erinnert das Land an die Ankunft der britischen Ersten Flotte 1788, mit der die europäische Besiedlung des fünften Kontinents begann.* (*dpa*, Issue 19, 2017)
[On "Australia Day" (Australia Day), the country commemorates the arrival of the British First Fleet in 1788, with which the European colonization of the fifth continent began].

Multiple types of flagging may appear on lexical items that are repeated within the same article. For example, there are three tokens of ≪*Box Jellyfish*≫, a highly venomous jellyfish (*chironex fleckeri*) found in the tropical waters of Australia, appearing in the same article. When it is introduced in the article headline, it is marked with quotation marks and a definition:

(23)   ≪*Box Jellyfish*≫—*Hochgiftige Qualle tötet Jugendlichen in Australien* (*dpa*, Issue 09, 2021)
["Box Jellyfish" — Highly venomous jellyfish kills boy in Australia].

only to be translated in the first sentence of the article as the unflagged *hochgiftige Würfelqualle*:

(24)   *Im Norden Australiens ist ein Jugendlicher durch den Stich einer hochgiftigen Würfelqualle ums Leben gekommen.* (*dpa*, Issue 09, 2021)
[In northern Australia, a youth was killed by a highly poisonous box jellyfish].

The third time the sea creature is mentioned, it is followed with *Seewespe* "see wasp," a translation of an alternative, but less common, English name for box jellyfish, in parentheses:

(25)   *Der 17-Jährige sei vor zehn Tagen beim Schwimmen an der Landzunge Cape York von den Tentakeln eines sogenannten Box Jellyfish (Seewespe) getroffen worden.* (*dpa*, Issue 09, 2021)
[The 17-year-old was touched by the tentacles of a box jellyfish (sea wasp) 10 days ago while swimming on the Cape York headland].

No longer requiring explanation or translation the final time it is mentioned, it remains unflagged inside a direct quote:

(26)   ≪*Wir sehen in unseren Gewässern sowohl Box Jellyfish als auch andere Quallenarten, die das Irukandji-Syndrom verursachen≫, warnten die Behörden.* (*dpa*, Issue 09, 2021)
[We see both box jellyfish and other species of jellyfish that cause Irukandji syndrome in our waters," the authorities warned].

The above examples show that flagging of Australia-related anglicisms is prevalent in articles written by journalists for the *dpa*. Analyzing the examples of flagging in the corpus reveals that a single word may appear in one article: (a) within quotation marks, (b) within quotation marks and in translated form, and (c) within quotation marks, in translated form, and accompanied by an explanation. The inclusion of flagged lexical units and codeswitches in the *dpa* articles may have the effect of placing local readers outside the English-speaking mainstream, as if they were peering in on a culture and society that they are not familiar with or part of, while for international readers, these flagging devices may provide local color and, in some cases, may assist in clarification of unfamiliar terms.

CONCLUSION

While the tradition for anglicism research has often been to use the print media published for German speakers living in German-speaking countries as a data source, this paper has applied a different approach by exploring the way that Australia, its icons, culture, institutions, and way of life have been portrayed using anglicisms in an Australian-published German-language newspaper.

It appears that a complex situation concerning the target readership influences the way Australia is portrayed in *Die Woche*, particularly when regarding the frequency of unadapted loans, loan translations, and loan renditions used to represent the broad themes of place and society. Because the newspaper under consideration here is published in Australia, there would be the assumption that its target audience of German speakers

living there is quite familiar with Australian English and is also acquainted with Australian concepts, geography, and society, etc. This, then suggests that there should be little need for loan translations or loan renditions, or even flagging of English lexical items, meaning that the number of unadapted anglicisms appearing within the corpus would be higher than attested. Nevertheless, only slightly more than half of all tokens in the dataset are in their original English form. While this oscillation between unadapted forms and adapted forms (including loan translations and loan renditions) adds lexical variety to the articles, there was a surprisingly high incidence of othering via the use of flagging devices on these unadapted lexical items. This has the effect of drawing the readers' attention to these items and marks them as being outside the recipient language.

Many articles within the newspaper are attributed to the *dpa* which reaches an international readership beyond the newspaper itself, including monolingual speakers of German in Europe. This would account for the relatively high number of loan translations and flagged lexical units and phrases than that which would be expected in a local publication for German speakers living in Australia. This would also explain the frequent appearance of the lexeme *Aborigine*, and its various hybridized forms, which has undergone semantic shift to become taboo in Australian English. While the inclusion of unadapted Australia-related anglicisms provides local color imagery, it risks a lack of understanding on the part of readers abroad who may be less familiar with English lexical items denoting place names, geographical locations, etc., within Australia. This, then, appears to create the need for many of the loan translations and/or flagging devices seen in the dataset.

The result of this, at times disjointed, combination of articles produced by local reporters for a local audience and articles produced for an international audience can sometimes lead to an othering effect upon the local reader, removing him or her from mainstream English-speaking Australia society and positioning him or her as an outsider looking in. Regarding this paper, it unfortunately leads to an unclear picture about the actual use of Australia-related anglicisms in German as used in Australia. To overcome this, future research in this area would benefit from using a dataset derived from a larger corpus, while at the same time, narrowing the focus by including only those articles specifically intended for a local audience written by local journalists. This would then allow for a comparison with a similarly large corpus sourced from the German print media, to give a better indication of how anglicisms are used in Australian German to represent concepts, phenomena, and cultural and societal artifacts related to Australia.

DATA AVAILABILITY STATEMENT

Non-copyrighted data supporting the conclusions of this article will be made available by the authors, without undue reservation. The copyrighted data analyzed in this study was obtained from https://www.diewoche.com.au. Requests to access the datasets should be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

REFERENCES

"bush, 1" (2004). *In Australian Oxford Dictionary*. Oxford University Press. Available online at: https://www-oxfordreference-com.ezproxy.newcastle.edu.au/view/10.1093/acref/9780195517965.001.0001/m-en_au-msdict-00001-0007658 (accessed February 27, 2022).

"bush, n. 1" (2021). *OED Online*. Oxford University Press. Available online at: https://www.oed.com/view/Entry/25179?rskey=Mm9NHe&result=1&isAdvanced=false (accessed October 29, 2021).

Adler, M. (2004). *Form und Häufigkeit Der Verwendung von Anglizismen in Deutschen und Schwedischen Massenmedien*. (Dissertation), Friedrich-Schiller-Universität, Jena (Germany).

Anthony, L. (2021). *"AntConc" (Version 4.0.0)*. Tokyo: Waseda University.

Burmasova, S. (2010). *Empirische Untersuchungen der Anglizismen im Deutschen am Material der Zeitung Die Welt (Jahrgänge 1994 und 2004)*. Bamberg: University of Bamberg Press. doi: 10.26530/OAPEN_368157

Carstensen, B. (1965). *Englische Einflüsse auf die Deutsche Sprache nach 1945*. Heidelberg: Carl Winter Universitätsverlag.

Carstensen, B., and Busse, U. (2001). *Anglizismen-Wörterbuch. Der Einfluß des Englischen auf den Deutschen Wortschatz nach 1945*. Berlin: Walter de Gruyter.

Clyne, M. (2003). *Dynamics of Language Contact*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511606526

Die Woche Australien (2017). Available online at: https://www.diewoche.com.au/geschichte-der-zeitung/ (accessed August 10, 2021).

Donalies, E. (1992). Hippes hopping und toughe trendies. über "(neu)modische", noch nicht kodifizierte anglizismen in deutschsprachigen female-yuppie-zeitschriften. *Deutsche Sprache* 2, 97–110.

Fiedler, S. (2012). "Der elefant im raum. The influence of English on German phraseology," in *The Anglicization of the European Lexis*, eds C. Furiassi, V. Pulcini, and F.R. Gonzaléz (Amsterdam: John Benjamins), 239–259. doi: 10.1075/z.174.16fie

Gardner-Chloros, P. (2010). "Contact and code-switching," in *The Handbook of Language Contact*, ed R. Hickey (Somerset: John Wiley & Sons), 188–207. doi: 10.1002/9781444318159.ch9

Görlach, M. (2001). *A Dictionary of European Anglicisms: A Usage Dictionary of Anglicisms in Sixteen European Languages*. Oxford: Oxford University Press.

Gottlieb, H. (2005). "Anglicisms and translation," in *In and Out of English: For Better or Worse*, eds G. Anderman, and M. Rogers (Clevedon: Multilingual Matters), 161–184. doi: 10.21832/9781853597893-014

Gottlieb, H. (2020). *Echoes of English: Anglicisms in Minor Speech Communities - with Special Focus on Danish and Afrikaans.* Berlin: Peter Lang. doi: 10.3726/b15937

Gottlieb, H., Andersen, G., Busse, U., Mańczak-Wohlfeld, E., Peterson, E., and Pulcini, V. (2018). Introducing and developing GLAD - the global anglicism database network. *ESSE Messen.* 27, 4–19.

Hunt, J. W. (2019). Lexical hybridization of English and German elements: a comparison between spoken German and the language of the German newsmagazine Der Spiegel. *Stud. Linguist. Univer. Iagellon. Cracovien.* 136, 107–120. doi: 10.4467/20834624SL.19.010.10605

Knospe, S. (2014). *Entlehnungen oder Codeswitching? Sprachmischungen mit den Englischen im Deutschen Printjournalismus.* Frankfurt am Main: Peter Lang.

Korff, J. (2020). *What is the Correct Term for Aboriginal People?.* Available online at: https://www.creativespirits.info/aboriginalculture/people/how-to-name-aboriginal-people (accessed September 4, 2021).

Langer, N. (1996). *Anglizismen in der deutschen Pressesprache. Untersucht am Beispiel von den Wirtschaftsmagazinen 'CAPITAL' und 'DM'.* Wettenberg: VVB Laufersweiler.

Matras, Y. (2009). *Language Contact.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511809873

Maude, A. (2018). Geography and powerful knowledge: a contribution to the debate. *Int. Res. Geogr. Environ. Educ.* 27, 179–190. doi: 10.1080/10382046.2017.1320899

Myers-Scotton, C. (1992). Comparing codeswitching and borrowing. *J. Multiling. Multicult. Dev.* 13, 19–39. doi: 10.1080/01434632.1992.9994481

Onysko, A. (2007). *Anglicisms in German: Borrowing, Lexical Productivity, and Written Codeswitching.* Berlin: Walter de Gruyter. doi: 10.1515/9783110912173

Onysko, A. (2019). "Processes of language contact in english influence on German," in *English in the German-Speaking World*, ed R. Hickey (Cambridge: Cambridge University Press), 185–207. doi: 10.1017/9781108768924.010

Onysko, A., and Winter-Froemel, E. (2011). Necessary loans - luxury loans? Exploring the pragmatic dimension of borrowing. *J. Pragmat.* 43, 1550–1567. doi: 10.1016/j.pragma.2010.12.004

Pflanz, M-L. (2014). Emprunt lexical: existe-t-il une typologie de la phase néologique? *Neologica.* 8, 157–183. doi: 10.15122/isbn.978-2-8124-2999-6.p.0157

Plümer, N. (2000). *Anglizismus - Purismus - Sprachliche Identität: Eine Untersuchung zu den Anglizismen in der Deutschen und Französischen Mediensprache.* Frankfurt am Main: Peter Lang.

Poplack, S. (1993). "Variation theory and language contact," in *American Dialect Research: Celebrating the 100th Anniversary of the American Dialect Society, 1889–1989*, ed D. R. Preston (Philadelphia, PA: John Benjamins Publishing Company). doi: 10.1075/z.68.13pop

Poplack, S. (2018). *Borrowing: Loanwords in the Speech Community and in the Grammar.* New York, NY: Oxford University Press. doi: 10.1093/oso/9780190256388.003.0004

Pulcini, V., Furiassi, C., and Rodríguez González, F. (2012). "The lexical influence of english on European languages: from words to phraseology," in *The Anglicization of European Lexis*, eds C. Furiassi, V. Pulcini, and F. R. González (Amsterdam: John Benjamins), 1–24. doi: 10.1075/z.174.03pul

Saugera, V. (2012). How english-origin nouns (do not) pluralize in French. *Lingvist. Investigat. Int. J. Linguist. Lang. Resourc.* 35, 120–142. doi: 10.1075/li.35.1.05sau

Schaefer, S. J. (2021). English on air: novel anglicisms in German radio language. *Open Linguist.* 7, 569–593. doi: 10.1515/opli-2020-0174

Sharp, H. (2007). Swedish–English language mixing. *World Engl.* 26, 224–240. doi: 10.1111/j.1467-971X.2007.00503.x

Winford, D. (2003). *An Introduction to Contact Linguistics.* Oxford: Blackwell.

Winter-Froemel, E. (2021). "Appealing though nebulous? introducing the concepts of accessibility and usability to loanword research [Conference presentation]," *Language Contact in Times of Globalization 5 (LCTG 5)*. Klagenfurt: The University of Klagenfurt.

Yang, W. (1990). *Anglizismen im Deutschen: Am Beispiel Des Nachrichtenmagazins Der Spiegel.* Tübingen: Max Niemeyer. doi: 10.1515/9783111676159

Zenner, E., Speelman, D., and Geeraerts, D. (2012). Cognitive Sociolinguistics meets loanword research: measuring variation in the success of anglicisms in Dutch. *Cogn. Linguist.* 23, 749–792. doi: 10.1515/cog-2012-0023

Zindler, H. (1959). *Anglizismen in Der deutschen Presse nach 1945.* (Dissertation), Universität Kiel, Kiel (Germany).

Check for updates

Irish English as a World English

John Kirk*

Department of English and American Studies, University of Vienna, Vienna, Austria

The objective of this chapter is to apply a synthesis of models for world Englishes to Irish English, for which data from ICE-corpora, the *Handbook of Varieties of English*, the *Mouton Atlas*, eWAVE, and the GloWbE corpus are used, and to which many approaches are applied: geo-political, dynamic-model, corpus, statistical, and multi-functional factor analysis. The chapter's relevance lies in its bringing together wide-ranging and seemingly disparate material through the lens of Irish English as a common denominator. The chapter shows the difficulties of applying top-down models to empirical corpus data and concurs that the gap is ultimately unbridgeable. Nevertheless, in so far as status distinctions can be inferred, this chapter also concurs with the many different findings that, as a standardized variety, despite a historical legacy of contact, Irish English is indeed an L1.

Keywords: Inner and Outer Circle, dynamic model, contact, koinéisation, functionality, multi-dimensional factor analysis, universals

INTRODUCTION

This chapter will explore the relationship between Irish English and world English(es). It will cover many approaches and raise many issues. It will show the need for the notion of "Irish English" to be recurrently clarified—especially the distinction between the vernacular variety and the standardized variety; and it will acknowledge the importance of a critical awareness between the two. The chapter is also mindful that, because of the presence of both Hiberno-English features as well as features from Scots, as shown in standardized data, the wider reality of educated public discourse in Ireland is never fully standardized.

Three research questions arise in this chapter: How has Irish English been characterized as an English among world Englishes given the range of approaches used in different investigations? What status does Irish English have as a world English as a result of the particular approach used? What overall conclusion about the status of Irish English as a world English may be drawn?

Several possibilities regarding status within Ireland and between Ireland and Great Britain have been proposed. This chapter will consider the status which has been accredited to Irish English in those top-down models of varieties of English around the world, especially those by McArthur (1987), Görlach (1990), Mesthrie and Bhatt (2008), Mair (2013), Siemund (2013), and Onysko (2016), among a few others on the one hand, and on the other by the quite different approach taken by Schneider (2007). Schneider (2007) Dynamic Model will then receive a short summary of an interpretation by Ronan (2020). The general challenge for all models of world Englishes is their demonstration and vindication of hypotheses on the basis of empirical corpus data. The few attempts that have been undertaken are reviewed by Hundt (2020), whose findings will be presented and whose conclusion that the gap would appear unbridgeable will be acknowledged.

The chapter will go on to recognize that there is no shortage of both standardized as well as non-standardized data (resources ICE, eWAVE, GloWbE, and so on, to be discussed below) for comparison and statistical evaluation. A further section will use the GloWbE Corpus to deal with the status of lexical items originating in Ireland, some of which will be shown to have remained there as Irishisms, whereas others have spread their wings and are now to be found in almost every variety around the world.

These global comparisons of individual items will lead into the final part of the chapter, which is concerned with various statistical attempts to assess and relate Irish English data to a great range of varieties worldwide, including pidgins and creoles. Whether ICE-corpora are used for data or the list of observation data behind eWAVE, whether multi-dimensional factor analysis is used on the standardized data, or whether multiple correspondence analysis is used on the reported observations of Hiberno-English data, the chapter will show that Irish English becomes repeatedly vindicated as an L1 or (in the Schneider, 2007 model) a stage-5 differentiated, variety, *tout court*, in answer to the third research question.

To these ends, such a synthesis arises from a combination of the relevant research literature that has so far been published as well as a fresh application of some of these approaches to Irish English data. The approach taken in this chapter is also informed by the author's knowledge and experience both as a resident in Northern Ireland since 1983, as a dialectologist of Irish English, and as a corpus linguist, particularly of Irish English and increasingly of world Englishes.

THE EVOLUTION OF CONCEIVING IRISH ENGLISH AS A WORLD ENGLISH

This first section deals with how Irish English has been conceived within the various frameworks for the study of world Englishes which have been proposed. It was during the 1990s and early 2000s that there emerged a new paradigm for conceptualizing the status of national varieties of English: that of World Englishes.[1] The investigation of this paradigm was facilitated by the growing provision of national components of English for the International Corpus of English (ICE) corpus (cf. Greenbaum, 1996), particularly those of L2 varieties, where English was becoming shaped by mother tongue speakers absorbing features of indigenous languages and where speakers of those languages were gradually acquiring English, by immersion, bringing many of their own mother-tongue features with them, or by instruction, usually in a predefined norm of standardized English. A central

question which arose during a review of the ICE project (Kirk and Nelson, 2018) was how far the models used for the delineation of world Englishes were adequate to deal with such multilingual environments, or whether they were becoming outmoded—such as for instance, the Quirkian notion of a "monochrome international standard language" with only local deviations (Quirk et al., 1985, p. 6).

Languages and their related offspring have traditionally been conceived of as trees, so when that tradition is superimposed on a map of the world, as Strevens (1980) did, it shows two main roots of British English and American English from which all other branches are deemed to stem.

Not surprisingly, as **Figure 1** shows, the line to Ireland is the shortest. The lines are drawn on a geo-political basis and identify countries around the world where English (of some sort) is to be found. The schema simply denotes Irish English as a descendant of British English.

Another image for describing languages is that of the wheel—of the hub and the rim, with the spokes in between.

In **Figures 2**, **3**, the hub is shown to be the part which is invariant and which holds all the other parts together—labeled as "World Standard English" by McArthur (1987) and "International English" by Görlach (1990), each not unreminiscent of Quirk's "monochrome international standard". Indeed Mesthrie and Bhatt (2008) comment that McArthur's "World Standard English" is "obviously an idealization" (p. 27), perhaps best represented as "written international English". For McArthur, "World Standard English" is a hub surrounded by a set of spokes, each labeled for a geographical area—with separate spokes for "British English" and "Irish English". The space in between is labeled as "British and Irish Standard English", as if a single entity. This view would appear to perpetuate the then prevailing view of Irish English as simply a matter of extra-territorial expansion, under exonormative (i.e., British) control.

In Görlach (1990) circle, the spokes are not labeled but, as the concentric circles radiate out from the hub, the set of spaces in the first circle are labeled as "regional national standards", the second as "sub-regional semi-standards", and the third and outermost circle labeled as "dialects", "ethnic semi-standards" or "non-standards". Thus it is as a "regional national standard" that British English is labeled, and Irish English as a mere "sub-regional semi-standard", albeit in the company of Scottish English, Welsh English and even English English. Elsewhere, Görlach (1995) laments that many words of Irish English do not appear in standard reference dictionaries and may not be known to people even with a good knowledge of English. Görlach (1995) challenge is addressed by Kirk and Kallen (2011) by looking at the vocabulary contained in ICE-Ireland (Kallen and Kirk, 2008). They show that there are many cultural references which are well known in each part of Ireland, of which people have clear awareness in their heads, which are used widely in broadcast media, but which are neither shared with the other part of Ireland nor with anywhere else. Examples include *decommission(ing), peace process/dividend,* and *maracycle* (cf. Kirk and Kallen, 2011). Such terms involve familiar, everyday words used in specialized senses or with a meaning somehow altered and thereby expressive of a particular cultural value.

---

[1]Since those times, the interest in World Engslishes has grown exponentially, not least because of the increasing availability of ICE corpora. Publications range from handbooks (Filppula et al., 2017; Schreier et al., 2020; Kirkpatrick, 2021), numerous research monologs (Schneider, 2007; Mesthrie and Bhatt, 2008; Siemund, 2013), collections of research papers (e.g., Suárez-Gómez and Seoane, 2015; Seoane and Suárez-Gómez, 2016; Dehors, 2018; Ling Low and Pakir, 2018; De Costa et al., 2019; Buschfeld and Kautsch, 2020; Meierkord and Schneider, 2021; Onysko, 2021; Saraceni, 2021; Siemund and Leimgruber, 2021) and finally textbooks (Kirkpatrick, 2007; Siemund et al., 2012; Jenkins, 2014; Melchers et al., 2019; Lange and Leuckert, 2020; Schneider, 2020).
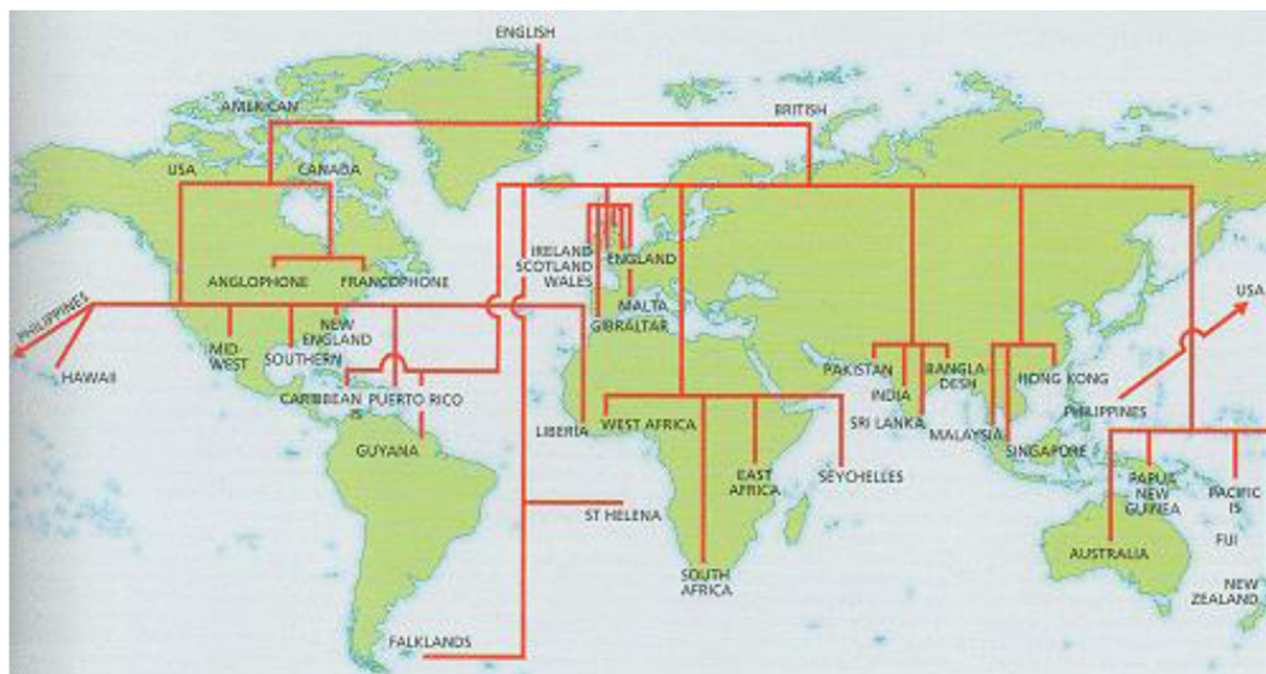
FIGURE 1 | Strevens's Geo-political Map of world Englishes (from Crystal, 1995, p. 107, redrawn from Strevens, 1980).

None of these words or expressions belongs to traditional dialect, and only some to the Irish language. Rather, Kirk and Kallen develop a cultural hypothesis which came to be substantiated by local semantic extensions or refunctionalisations in English, from the result of which emerges a cultural vocabulary which is finely-hewn for each polity, Northern Ireland and the Republic of Ireland. Nevertheless, this vocabulary remains still relatively unfathomable to outsiders, so that a glossary for ICE-Ireland would not go amiss.

Shortly before the McArthur (1987) and Görlach (1990) models were published, a set of very different circles was proposed by Kachru (1982), as shown in **Figure 4A**.

Kachru extends Strevens's geo-political varieties to combine national varieties into groups depending on the status of English there: as a national language (ENL) or mother tongue (EMT), as an official additional or second language (ESL or EAL), or as a foreign language (EFL). These groupings comprise three non-overlapping circles respectively: an inner circle, an outer circle, and an expanding circle, which have been drawn by others in various ways. The outer circle comprises countries to which English had been taken during period of colonization such as India, Kenya, or Singapore. The expanding circle comprises countries where English has had no internal or institutional status such as most countries in Europe but also countries such as China or Brazil. To each of the three circles Kachru (1982) added a further designation: ENL-countries were "norm-providing", ESL countries were "norm-taking", and EFL countries were "norm-dependent", as shown in **Figure 4B**.

Although Kachru's model was concerned with postcolonial Englishes of relatively recent times, it was not supposed to be concerned with "old" colonies such as Ireland. Thus, after centuries of English use, Irish English was placed in the norm-providing, native/mother tongue category. That Irish English qualifies as a norm-provider may be adduced from its diasporic effects—in Newfoundland, in American English, and in emigration to the urban areas in Great Britain such as Glasgow, Liverpool or London. Although Kachru (1982) model has been much discussed and become hugely influential, as the present focus is on Irish English, many of the specific criticisms need not concern us here. However, Kachru (1982) model has recently been vindicated by Axel Bohmann (2020) in a multi-dimensional factor analysis of some 10 ICE-corpora, as discussed below.

Kachru's model shifted the focus away from nation states to different categories of speakers and users of the language, including learners. For Modiano (1999), for instance, among learners it's as much a question of proficiency in the language as it is with native speakers.

Although Ireland doesn't figure explicitly in Modiano's model (see **Figure 5**), the distinction around language functionality which he makes can be plausibly conferred on the Irish situation: between the low registers in the vernacular (often referred to as Hiberno-English) and confined to Ireland, and standardized English, acquired largely through education and exposure to situational need, which takes care of the higher registers and international communication.

Reconceptualisation was taken further by Mesthrie and Bhatt (2008) in what they came to call "the English Language
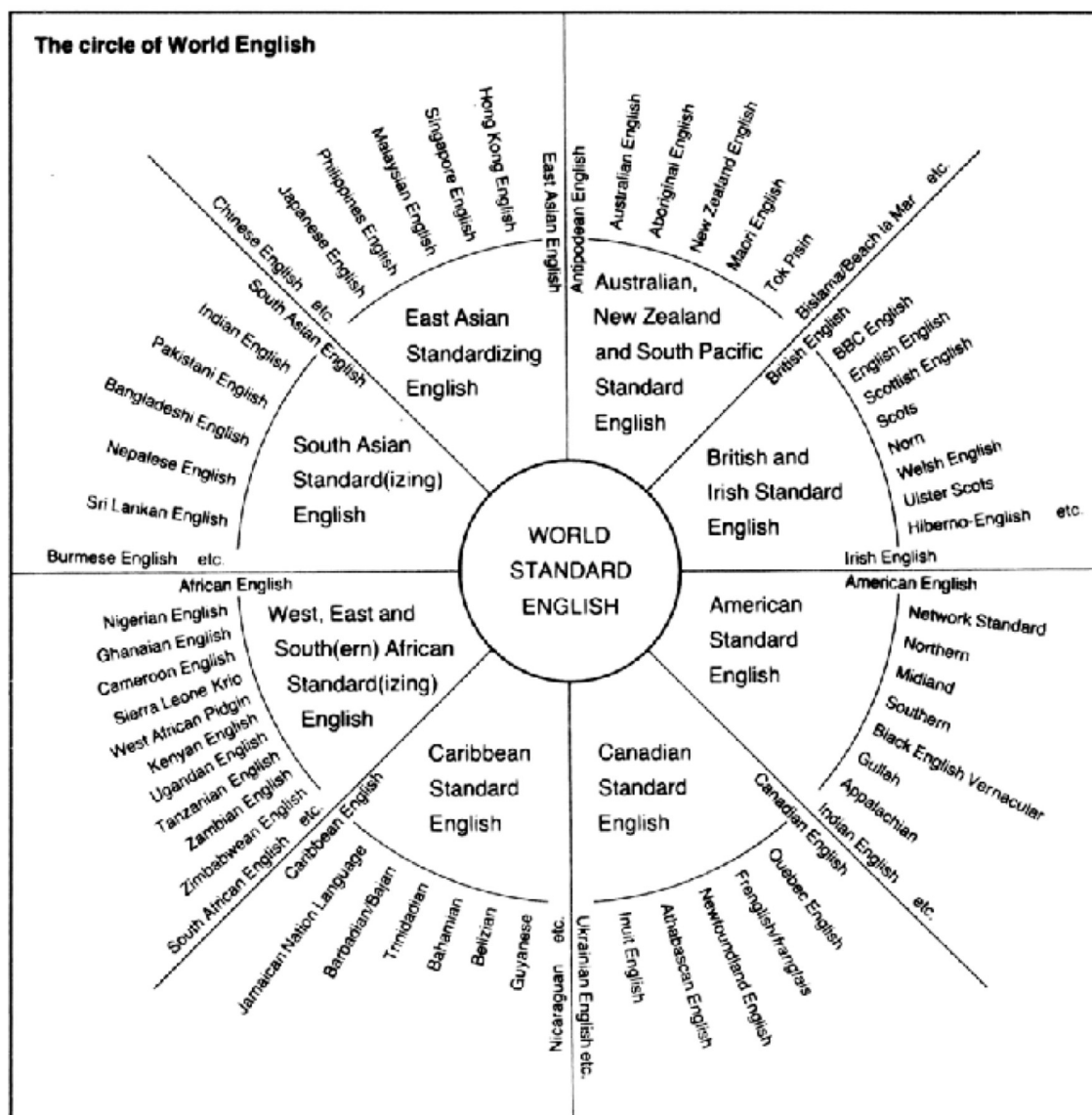
FIGURE 2 | McArthur's "The Circle of world English" (from McArthur, 1987, p. 11).

Complex", with no fewer than 12 subtypes.[2] Of those, they claim that Hiberno-English, once a "language-shift variety",[3] i.e. a variety which "has at some crucial stage of its development involved adult and child L1 and second-language (L2) speakers [as] one speech community [...] with L1 and L2 speakers of the dialect closely interacting with each other" (p. 6), is now "probably best classified as a social dialect" (p. 6) bound up in a community with only L1 speakers. Nevertheless, despite its comparability with other L1 dialects within Britain, and its role

in turn as a superstrate variety in certain later colonial situations, Mesthrie and Bhatt (2008) revert to its origins as "a language shift variety" citing Hickey's observation that "the majority of the Irish acquired English in an unguided manner as adults" (p. 44). Without this significant contact with L1 speakers, Mesthrie and Bhatt (2008) claim that Irish English as an L2 came to result in and stabilize as an L1, and thus, because of its linguistic features similar to other colonial Englishes, Irish English is, in effect, a "new English" or a quasi-outer circle variety. This rather inductive view would seem be somewhat out of keeping with more empirical claims and findings discussed below.

A further model (**Figure 6**) along geo-political lines is the "World System of Englishes", put forward by Mair (2013), wherein circles are replaced by a hierarchical list, with a hub at
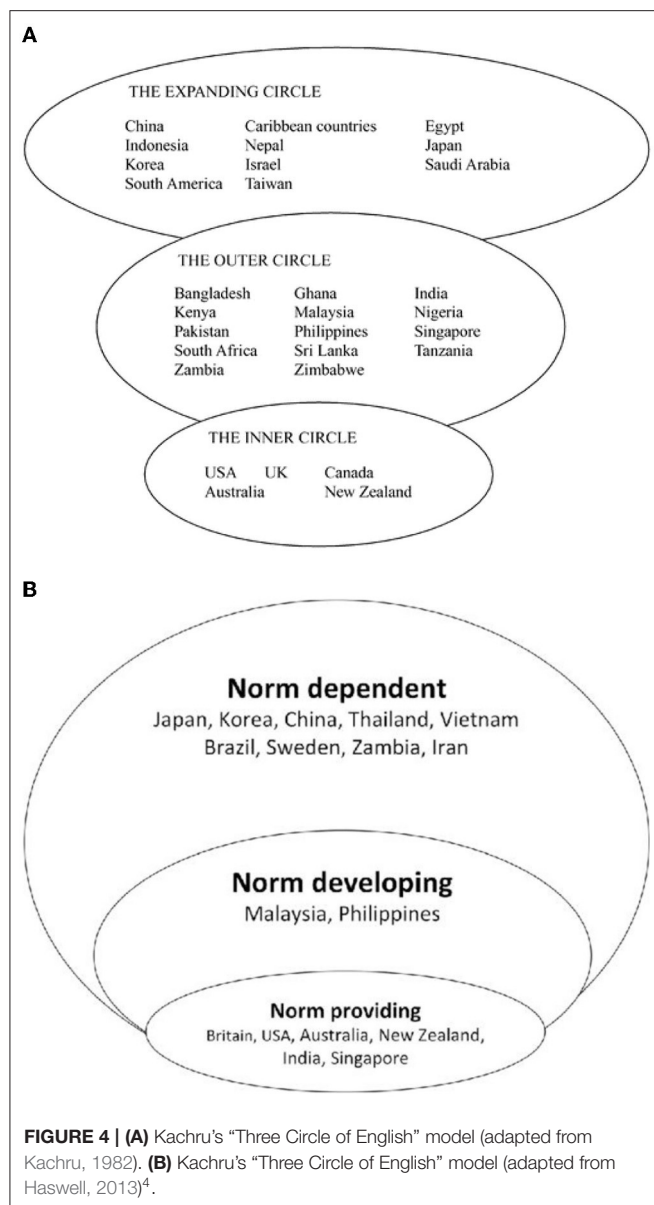
---

[2]These 12 subtypes are "metropolitan standards", "colonial standards", "regional dialects", "social dialects", "pidgin Englishes", "creole Englishes", "English as a second language", "English as a foreign language", "immigrant Englishes", "language-shift Englishes", "jargon Englishes", and "hybrid Englishes".
[3]For Siemund (2013), too, Irish English ranks as a shift variety as a result of contact.

FIGURE 3 | Görlach's "Circle of International English" (from Görlach, 1990).

the top of this list, from which all other levels thence descend. Irish English is ranked as a "central variety", perched along with (explicitly) "Scottish Standard English" between super-central varieties" (British English) and "peripheral varieties". An innovation in Mair's model is the inclusion and linking at each level of a separate hierarchy of non-standard varieties, ranging from the same set of levels. In terms of this model, if Irish standard(ised) English ranks as "central", Hiberno-English ranks merely as "peripheral".

The World System of Englishes entails a cascading-down approach: super-central varieties will influence central varieties but not *vice versa*—thus British English on Irish English but not *vice versa*. As Mair comments: "there will be more British terms spreading into Irish usage than in the other direction." (Mair, 2013, p. 262). It is doubtful that Irish English will cascade-down to influence "peripheral varieties", unless perhaps, because of nineteenth-century emigration, on urban varieties in England such as Liverpool or Newcastle-upon-Tyne. But yet again, it depends on what is being meant by Irish English: simply vernacular Hiberno-English, whether northern or southern? And moreover what is meant by influence, or who the influencer is.

FIGURE 4 | (A) Kachru's "Three Circle of English" model (adapted from Kachru, 1982). (B) Kachru's "Three Circle of English" model (adapted from Haswell, 2013)[4].

A more recent top-down classificatory model—the "Language Contact Typology of world Englishes", by Onysko (2016) reproduced in **Figure 7**—is based on the premise that all variation arises from contact. Here, the early foundation of English in Ireland may be classified as an "English in multilingual constellation".

In his model, Onysko (2016) applies Van Coetsem (1988, 2000) distinctions between a "source language" (in this case Irish) and a "recipient language" (in this case English). Onysko's claim is that, if the source language is dominant in the contact situation, its impact on the recipient language will go beyond basic lexical influence (assumed to be borrowing) and be significant. In

this contact situation, speakers of the recipient language (English) will thereby feel pressure to acquire the source language (Irish) and adopt many of its features from different levels. The corollary claim is that, if the recipient language (English) is dominant in the contact situation, it will remain dominant and absorb only lexical features from Irish. Thus, Onysko contends, it was this essentially parallelism in the bi-lingual contact situation which laid the foundations for what eventually led to the wholesale language shift from Irish to an Irish form of English, stretching from the late 18th to the early 20th centuries. From the view of the contact model (**Figure 7**), the emergence and early foundation of Irish English fits the general scenario of "Englishes in multilingual constellations" whereby continuing source language agentivity turned into a scenario of large scale language shift.[5]

Either way, many Irish words found their way into Irish English. Although these words are considered as "loanword borrowings", Kallen (1996) has urged that they are probably better understood as "apports". For Kallen, apports are "Irish words retained in the English language after a community switch away from Irish" (Kallen, 1996, p. 109, cf. also Kirk, in press). The notion suggests the source-language activity of ""survival" and the "carrying over" of elements in the Irish lexicon [which speakers of Irish felt they couldn't give up]— rather than borrowing by English [and it] suits much of the development of Irish English, especially during the major period of language shift" (Kallen, 2013, p. 132).

After the completion of the shift for large parts of the population, Irish English can be considered as developing further—largely as a "Koiné English" (i.e., as a result of dialect contact between the emerging variety of Irishised English and other varieties of English).

More recently, a further contact development was to develop: spoken vernacular English in Ireland has come to find itself in contact with other dialects and varieties of English, most notably the standardized variety, but also British regional varieties, reinforcing further the classification of Irish English as a "koiné". Koiné features from dialectal English are certainly shown to be present in southern Hiberno-English by Klemola (1994, 2002) and Anderwald (2002), and those from Scots in northern Hiberno-English (Kirk and Kallen, 2010)—all further evidence of source-dialect activity and of apporting rather than borrowing. Most recently, Irish English is facing a new Koiné -contact situation. whereby it is *Irish* English that is being acquired as their English by communities of migrants, most predominantly from Poland. Thus in time, Irish English stands to become subjected to modification though their learning efforts (cf. e.g., Migge, 2012; Nestor et al., 2012; Diskin and Levey, 2019; Corrigan, 2020)— a further instance of emerging bilingualism in a multilingual constellation, with Irish English as the target L2 language for L1 speakers of Polish.

Against these top-down, geo-political approaches at explaining the relationships between world Englishes, there has arisen an alternative paradigm with an altogether different

---

[4]Haswell (2013) presents other models of world Englishes, but as their focus is on the Asian rim, they will not be discussed further here.

[5]All the same, Irish multi-/bi-bilingual contact almost certainly continued after the language shift with prevailing code-switching, more likely among educated members of the middle-classes.
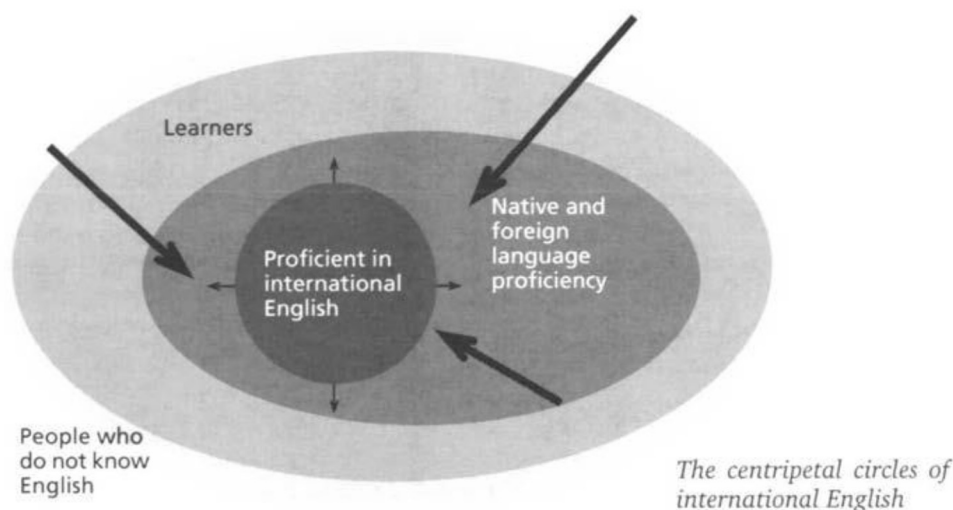
FIGURE 5 | Modiano's "Centripetal Circles of International English" (from Modiano, 1999, p. 25).


"World System of Standard and Non-Standard Englishes"

- **Hyper-central variety / "hub"** of the World System of Englishes: Standard American English
- **Super-central varieties:**
  (1) Standard: British English, Australian English, South African English,
  Nigerian English, Indian English, and a very small number of others
  (2) Non-standard: AAVE, Jamaican Creole, popular London, and a very small number of others
  (+ Domain-specific ELF uses: science, business, international law, etc.)
- **Central varieties:**
  (1) Standard: Irish English, Scottish (Standard) English, Jamaican English, Ghanaian English, Kenyan English,
  Sri Lankan English, Pakistani English,
  New Zealand English, and a small number of others
  (2) Non-standard: Northern English urban koinés, US Southern, and a small number of others
- **Peripheral varieties:**
  (1) Standard: Maltese English, St. Kitts English, Cameroonian English,
  Papua New Guinea English, and others
  (2) Non-standard: all traditional rurally based non-standard dialects, plus a large number of colonial
  varieties including pidgins and creoles

FIGURE 6 | Mair's "World System of Standard and Non-Standard Englishes" (from Mair, 2013, p. 264).

approach. This is the "Dynamic Model", advanced by Edgar Schneider, initially in an article (Schneider, 2003), later in more detail in a monograph (Schneider, 2007). The focus of this model is on the English of former (mostly British) colonies, and is "dynamic" because it sets out explanations for the evolution and development of the Englishes in these countries. There is a basic premise whereby Englishes developed through an evolutionary process which is uniform, and which can be

accounted for in separable stages. In such counties, two groups of speakers are always identifiable: firstly, the settlers (in the case of Ireland, English in the South; English and Scots in the North) who brought their mother tongue with them and who, over time, come to acquire features taken over from local languages native to the place; and, secondly, the indigenous, colonized people—the Irish—who begin to acquire the language of their new colonial masters. Initially, for the early phases, the English

FIGURE 7 | Onysko's "Language Contact Typology of world Englishes" (from Onysko, 2016, p. 213).
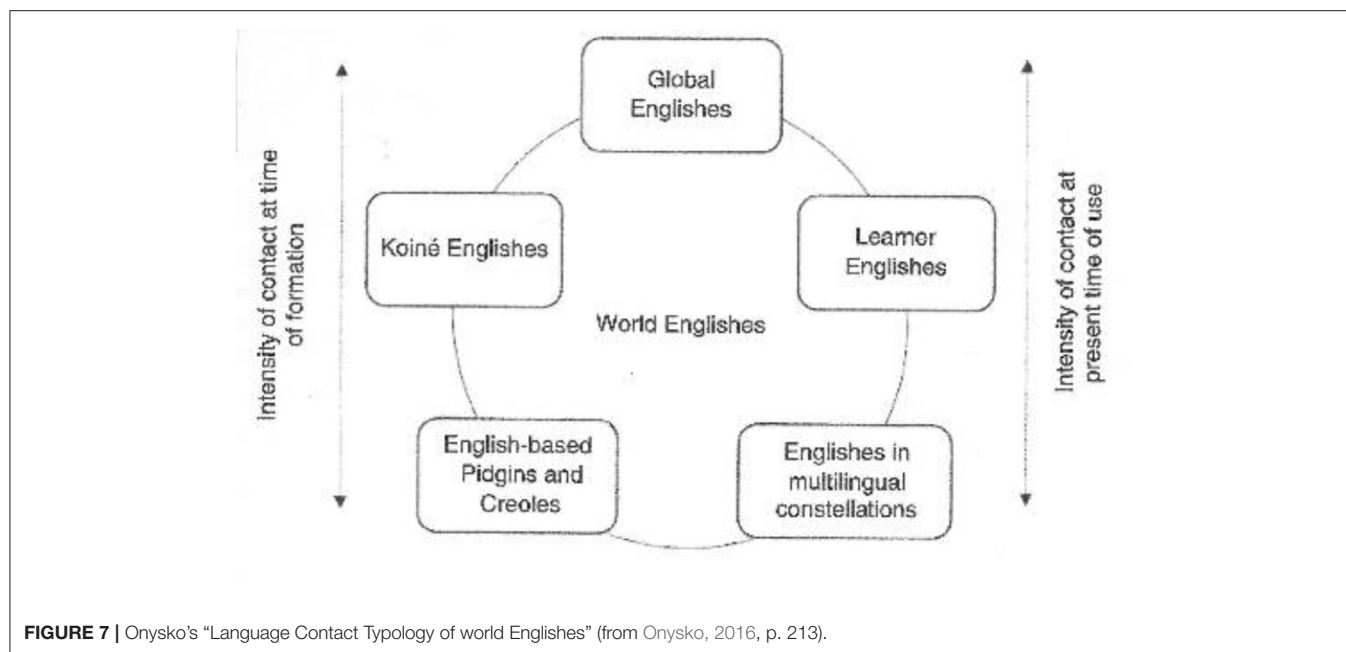
is based on that of the settlers and follows their (largely British) "norm-providing" or "exonormative" patterns (Schneider's second tier). Over time, and inter-generational transfer and inter-marriage, and in Ireland advanced by language shift, the English of each group comes to merge, with features becoming transferred from Irish and now shared in common and thus nativized (Schneider's third tier) (similar to Mesthrie and Bhatt's "language-shift" variety above). Over further time and deeper integration, features come to have local characteristics of their own or "endonormative" patterns (Schneider's fourth tier). Finally, there is full linguistic independence or "differentiation" (Schneider's fifth and final tier) or independent variety.[6]

So what of English in Ireland, Britain's oldest colony? A trenchant case for the development of English in Ireland in line with the Dynamic Model is made by Ronan (2020). Ronan recognizes that the model's foundation stage is not met—English arrived on the back of Anglo-Norman French, the speakers of which were the colonizers and in due course the power-holders, so that, for much of the later Middle Ages, the prestige languages of Ireland were French, Irish, and Latin. As Hickey (2007) and Ronan (2020) comment, English didn't really take off in Ireland until the Plantations of the North in the early 17th century and in the South later that century—what amounted to a second colonization. Large numbers of Gaelic population groups began then to shift to English. For Ronan, "it is in [...] examples of seventeenth-century Irish English that we find examples of exonormatively standardized English on the one hand, and on the other hand heavily contact-marked varieties [...] which could

be seen as typical markers of ongoing nativisation of the English language in Ireland." (p. 336) As time went on, Ronan argues, "a situation of diglossia started to emerge, in which English was increasingly used in all domains while Irish was relegated to daily life particularly in rural areas." (p. 338) Here, Ronan contends, "we can speak of a second nativisation process of the English language". She elaborates: "English made further inroads when a National School system was introduced in 1831, and the death blow was dealt to [Irish] by the depopulation of rural areas during the famine years in the 1840s (p. 338). Over time, the widespread Irish monolingualism turned into bilingualism and was replaced by English monolingualism. The growing importance of English, especially in the economy and in politics, led to a steady decrease of Irish speakers in all areas of the country. "Arguably", Ronan continues, "it is this spread of English as the first language of the education system and the increasing loss of the Irish contact language which caused the endonormative stabilization of the language" (p. 338), including the incorporation of many well-known features of Irish as discussed above which had become transferred.[7]

Schneider's final stage of differentiation is equated by Ronan with regional dialect differences (northern vs. southern, as mentioned above) and social differences which are ongoing (as, for instance, with the Dublin accent, and its supra-regional

---

[6]An example of a variety clearly differentiated from British English is American English, with its own spelling variants, its many differences of morpho-syntactic variation, and especially its long-standing tradition of lexical codification, all the result of demographic and social merger and, in terms of Onysko's categorisation, "koinéisation".

[7]To explain the development of English toward endonormative stabilization Ronan's case is supplemented by the adoption of the Dynamic Model's sequel: the Extra- and Intra-Territorial Forces Model (Buschfeld and Kautsch, 2017), Ronan argues: "In effect, we find two colonization phases, with the second one starting in the seventeenth century, 600 years after initial settlements. Extra-territorial and intra-territorial language policies after this second settlement were considerably stricter, and the use of English, as well as the suppression of the Irish language, were enforced rigorously. Intra-territorially, the necessity to submit to the English language and culture was also stronger from the seventeenth century on than after the initial, twelfth-century settlements" (Ronan, 2020, p. 341).

influences further afield; cf. Hickey, 2005). A further case for differentiation must surely be the distinctiveness of the vocabulary and its extensive codification in recent dictionaries: *A Dictionary of Hiberno-English* (Dolan, 2014), *Slanguage,* (Share, 2014), and the *Concise Ulster Dictionary* (Macafee, 1996) (cf. also Kirk, in press).

Schneider's model has been well-received, with some debate about the stage of development to which a variety is allocated, and about whether some nativised varieties have developed their own norms. Recent vindication comes the multi-dimensional factor analysis in Bohmann (2020), as discussed below.

Meanwhile, the Dynamic Model's applicability to non-colonial countries has been tested in investigations of European and Asian countries, with positive results, although, inevitably, the first two stages cannot apply. In such countries, where English has no settler population or institutional status, it has been acquired entirely through education with an inevitable focus on the standardized language (Edwards, 2016, p. 190).

Thus, to accommodate developments in common, there has arisen the notion of what Schneider (2014) calls "transnational attraction", a further dynamic concept intended to explain the influence of English on languages everywhere, no matter their status. Behind this notion lie notions of globalization, of instantaneous electronic-mediated communication, widespread global travel, affluence in westernized societies and, increasingly, migration into large urban centers of multi-ethnicity or what's become labeled as super-diversity, such as Singapore, Hong Kong and Dubai (cf. Siemund and Leimgruber, 2021). Of such social influences, Irish English has certainly not been exempt, with many multi-national companies such as Apple, Google and Microsoft setting up production facilities in Ireland, or the influx of particularly eastern and southern Europeans to Ireland.

To sum up this synthesis of proposals for conceptualizing the status of national varieties, we have thus seen that Irish English has been categorized internal-linguistically as: "standard" (McArthur, 1987), "sub-regional semi-standard" (Görlach, 1990), "inner circle"/"ENL"/"EMT" (Kachru, 1982), "contact", "central" (Mair, 2013), "koiné" (Oynsko, 2016) and "differentiated" (Schneider, 2007); also externally as: "postcolonial" (Schneider) and "new" (Mesthrie and Bhatt, 2008). Can they all be credible? The differences of focus in Ireland have inevitably split between high and low prestige languages in an earlier period (as well documented in Kallen, 2013, p. chs. 1 and 5; Kallen, 2017) and, in the modern period, between vernacular Hiberno-English and the standardized variety. Although features of the vernacular occur in the standardized variety only in small measures, the latter is no less Irish in their indexicality. The two extremes remain joined through the abiding distinctiveness of the vocabulary.

## THE EMPIRICAL COMPARISON OF IRISH ENGLISH WITH WORLD ENGLISHES

Top-down conceptualization has not been the only approach to considering Irish English within world Englishes. For there has been a further development of a further paradigm. With

the growing availability of ICE-corpora, Irish English has come to feature as a national "dialect" in inter-corpus comparisons. A great many studies have compared the data of ICE-Ireland with the data of other ICE-corpora. The recent textbook by Lange and Leuckert (2020) encourages inter-ICE-corpora comparisons with a set of five exercises (on phrasal verbs, the definite article, embedded questions, processes of compounding and suffixation, and adverbial intensifiers). Such studies are facilitated by the ICEonline software which at a stroke can compare up to 15 national components. A recent article by Hundt (2020) reviews the research which uses ICE-data to substantiate the claims of various WE models as outlined here. Section §22.4.1 of Hundt (2020) is entitled "Corpora as a Testing Bed for Models of world Englishes"–specifically Kachru's Three Circles Model, Schneider's Dynamic Model, and the Epicenter Hypothesis.[8] With respect to the Three Circles Model, Hundt comments: "numerical evidence on individual grammatical patterns does not necessarily support the distinction between ENL, ESL, and EFL varieties …" (Hundt, 2020, p. 515) and provides examples of those which do and those which do not.

For testing Schneider's Dynamic Model of the evolution of WEs for different degrees of nativisation, Schneider himself urges that " the most promising road to a possible detection of early traces of distinctive features is a principled comparison of performance data collected along similar lines, i.e. systematically elicited corpora" (Schneider, 2004, p. 227). To this end, the verbal complementation patterns intransitive, monotransitive, or ditransitive, and their lexical exponence are investigated in ICE-India, ICE-Hong Kong, ICE-Singapore along with ICE-GB in a range of articles by Mukherjee and Gries (2009) and Gries and Mukherjee (2010). Hundt reports favorably: "[…] Mukherjee and Gries (2009)" collostruction analysis reveals that a variety's divergence from BrE tallies well with its respective developmental stage […]" (Hundt, 2020, p. 516), so that if ICE-Ireland data were used, and with stage 5 differentiation, close or comparable similarity with ICE-GB data would be likely. Using the same corpora, however, in an investigation of n-grams and their bundling, Hundt reports further: "Gries and Mukherjee (2010) study of lexical bundles did not provide additional evidence that the more advanced a variety is in the evolutionary cycle, the more it will have developed region-specific usage patterns" (Hundt, 2020, p. 516). Whereas these findings may hold for the Asian varieties which were investigated, there are two implications for Irish English: firstly, if applied to ICE-Ireland, distinctive patterns of lexical exponence in complement types may not be found either as the corpus is far too small for such purposes.[9] That notwithstanding, as differentiation in Schneider's model entails lexical distinctiveness, we know that such distinctiveness exists from the recent lexicographical records, as discussed above. From

---

[8]As Irish English can hardly be claimed to be an epicenter, the topic won't be developed further here.

[9]Recent online corpora such as GloWbE (with its sourcing of material solely from the Internet) or NOW (corpus of current newspaper material from the Internet) may well reflect more of Irish English's lexical distinctiveness. But see further below.

FIGURE 8 | eWAVE Atlas (from https://ewave-atlas.org/languages).
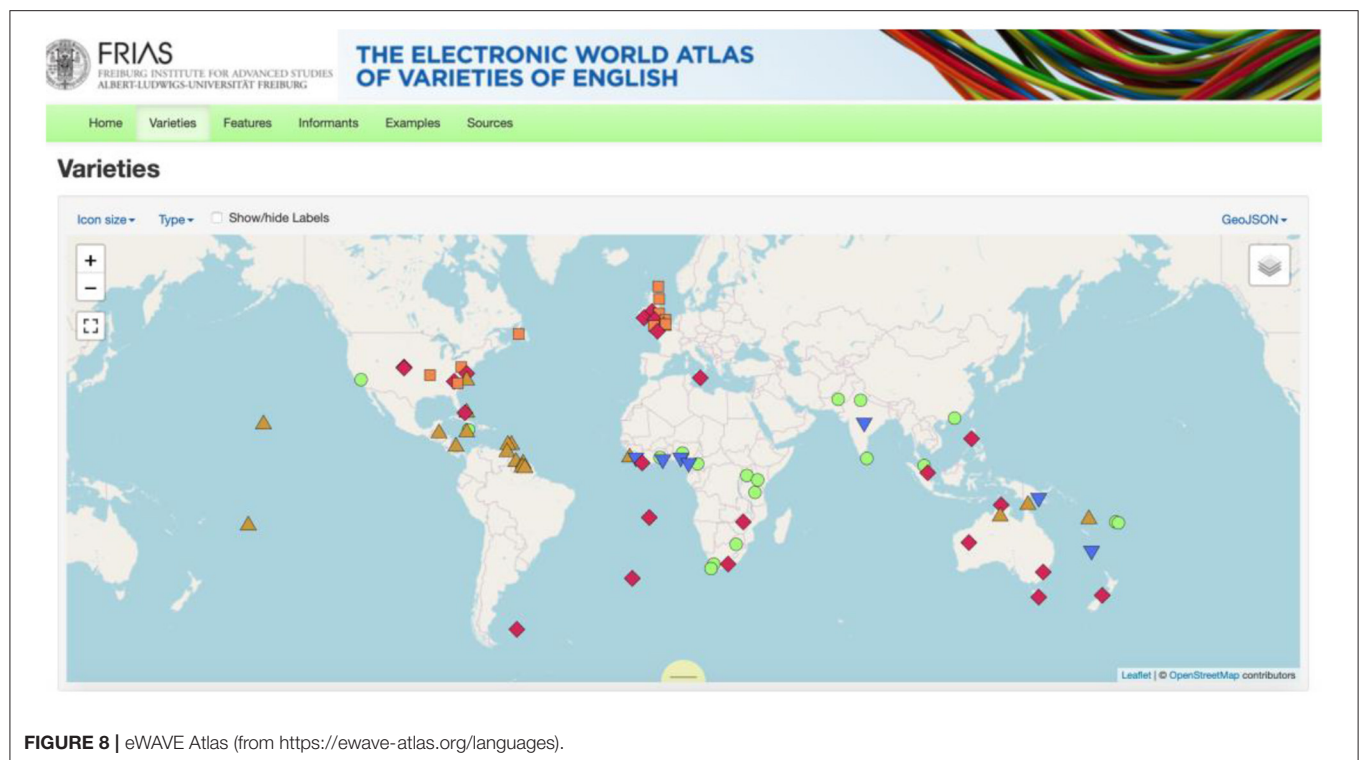
Hundt's comprehensive article, it would be fair to infer that there must be extremely few research publications which have used (or, moreover, can use) corpus data directly to substantiate the claims or implications of top-down, largely geo-political models of world Englishes. The inference gap appears to be simply unbridgeable—or at least so far.

In what other ways is it possible to relate and compare Irish English to other varieties of English. One is to use the *Electronic World Atlas of Varieties of English* (eWAVE),[10] compiled by Bernd Kortmann (see **Figure 8**). Here, information on maps and in lists for some 225 features in 77 countries is presented. It is a database, not any kind of corpus. The list of features was drawn up in the early 2000s by Kortmann on the basis of what was then known about non-standard features in varieties of English world-wide and collated into an organized list. A list of areas was compiled where coverage might be revealing and the inclusion of which was considered important. The upshot of all these data is that Irish English (covered by Markku Filppula) is categorized as a "high-contact L1 variety".

Of the 23 features given an A-rating ("feature is pervasive or obligatory") for Irish English in eWAVE, the rarest occur in only one other variety: the *after*-Perfect (feature 98) in Newfoundland, and the unsplit *for to* in infinitival purpose clauses (feature 202) in Tristan da Cunha; two other features occur in only three other varieties: the absolute use of reflexives (feature 15) in Manx English, Saramaccan English and Eastern Maroon Creole,

and forms or phrases for the second person singular pronoun other than *you* (feature 35) in Orkney and Shetland English, Gullah, and Saramaccan English. A further feature occurs in only four other varieties: the use of definite article where StE favors zero (feature 64) in Manx English, Indian English, Hong Kong English, basilectal Fiji English. For the far-flung distribution of some of these features, there is probably no single explanation. But in view of their rarity, it would not be inappropriate to consider them Irishisms (and the *after*-perfect usually is!). On the other hand, some of these "pervasive and obligatory" features are quite widely distributed: in some 60 (out the total of 77) countries, alternatives for the second person plural *you* pronoun (feature 34; see **Figure 9**) are listed for 63 countries, and the lack of inversion and elision of the auxiliary in questions (feature 229; see **Figure 10**) is attested in 67 countries.

The eWAVE maps are thus able to show at a glance how widespread a feature of Irish English actually is—with maps that not only show the A-D ratings[11] of the response but also the categorization of the variety (as an L1, L2 or creole type).

The online atlas derives from the earlier *Mouton World Atlas of Variation in English* (Kortmann and Lunkenheimer, 2013), where slightly different results are to be found. For Irish English, the *after*-perfect is the only A-rated feature. However, there are listed some 16 B-rated features (p. 695). In the *Mouton Atlas*, Irish English is again categorized as a "high-contact L1 variety",

---

[10]https://ewave-atlas.org (last accessed January 05, 2022). Cf. Kortmann et al. (2020).

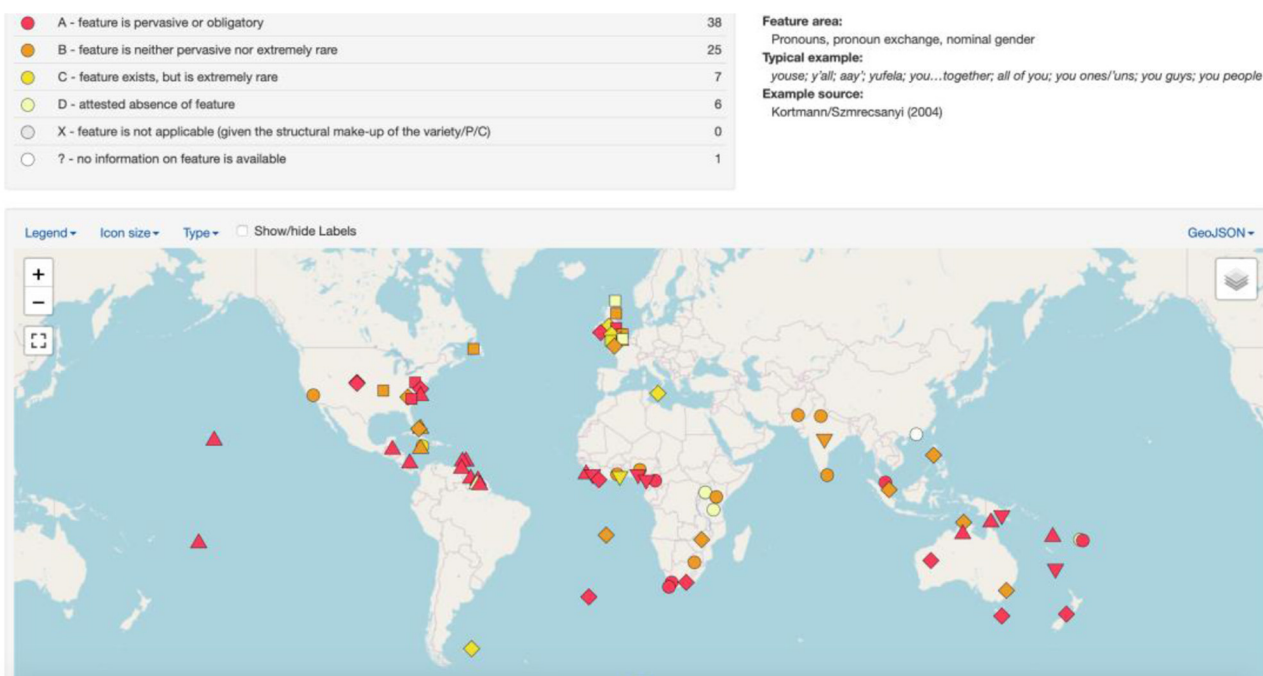[11]For B, a feature "is neither pervasive nor extremely rare"; for C, a feature "exists but is extremely rare"; for D, a feature is attested as "absent" (https://ewave-atlas. org/parameters/1#2/7.1/43.4).

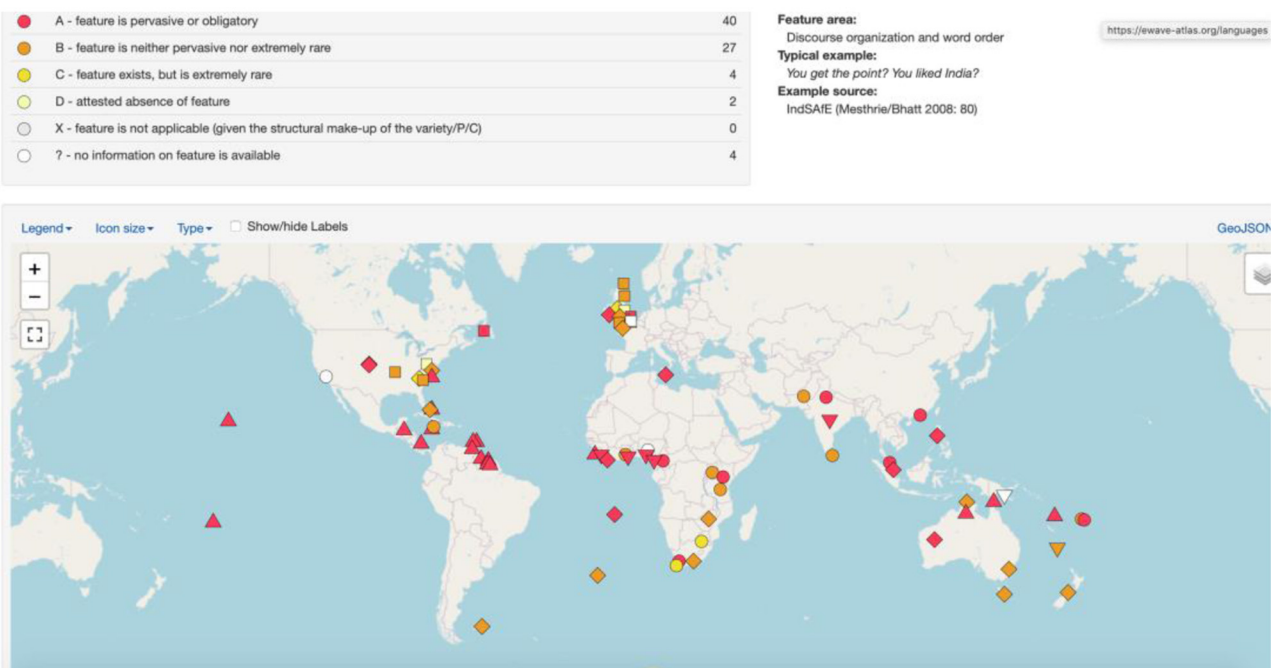**FIGURE 9 |** Screenshot of eWAVE Feature 34: "Forms or phrases for the second person plural pronoun other than *you*" *from https://ewave-atlas.org/parameters/34#2/7.0/7.7.

FIGURE 10 | Screenshot of eWAVE Feature 229 "No inversion/no auxiliaries in main clause yes/no questions" (from https://ewave-atlas.org/parameters/229#2/7.0/7.7).

one of five broad classes of variety types.[12] The *Mouton Atlas* presents various groupings such as the most frequent features across 28 of the 30 L1 varieties surveyed (p. 681–682). Whereas Irish speakers of English are aware of features they have and which seem to depart from the codified standard—and have responded well to acceptability judgements[13]—these speakers simply do not know how widespread within Ireland or how localized to Ireland any feature actually is. Without comparison, "it is not possible to establish any variety-exclusive and variety-preferential features of any (regional, socio-economic, ethnic, etc.) dialect", according to Barron and Schneider (2008, p. 429). It is not only to qualitative and quantitative comparisons of this kind that the *Mouton Atlas* lends itself but also to multi-factorial and typological analyses. For individual morpho-syntactic and discourse-pragmatic features, the *Mouton Atlas* and eWAVE stem from a database that has been rigorously annotated to enable multi-factorial and multi-dimensional analysis and interpretation. No doubt this has all been made possible by the nature of the data upon which they are based: sentences that were rated for their occurrence and acceptability.

The question then arises whether the automatic approach provided by eWAVE can be replicated on the basis of corpus data. As discussed above, Hundt (2020) finds that that very few corpus-based studies have satisfactorily substantiated the apparent differences claimed by world English models, whether those by Kachru or by Schneider. Nevertheless, there are two approaches using corpus data which go a long way toward operationalising instantaneous comparisons between Irish English and world English: the ICEonline software, already mentioned, and the GloWbE corpus.

It's well known that one-million word corpora such as ICE components are ideal for providing data for many morpho-syntactic or discourse-pragmatic investigations but much less for lexical phenomena. It was shown above how a key word approach to lexis, by looking at the incidence of vocabulary in the corpus as a whole, was able to substantiate cultural differences between the two parts of Ireland (cf. Kirk and Kallen, 2011). However investigating individual lexical items on their own proves unsatisfactory in corpora of that size because of the limited amount of data. Of the best known Irish words in English: *brogues, galore, gob, slew, slogan, whiskey* (cf. Phelan, 2016), ICE-Ireland yields only one instance of *galore*, and there are no occurrences in any other ICE component. A further category of words, claimed by Mullally (2017) to be Ireland's favorite words, includes: *hooligan. leprechauns, langers* and *eejits, donnybrooks* and *Tories* as well as the eponymous cases of *boycott*, each of which has its origins in Irish. Of this latter set, there are no occurrences in ICE-Ireland and only one occurrence of *Tory* in ICE-GB. More widely, there are 26 occurrences of *boycott* (shown by ICEonline). Happily, lexical enquiries on the basis of corpus

data can be undertaken with the use of the GloWbE corpus.[14] Not only can enquiries be executed online automatically but the results can be instantaneously displayed in a table which amounts to, if not a map of the world, as with eWAVE, a horizontal cartographical-like display of relative distributions with illustrations like bar graphs. GloWbE is a massive corpus of about 1.9 billion words of text from some twenty different countries, from the years 2012–2013. Its data are taken from the Internet, including blogs, so that it comprises a mixture of fairly informal material from both spoken-like and written-like texts, tending probably more toward the informal or colloquial end of a stylistic continuum (cf. Davies and Fuchs, 2015).

Of the above words which Ireland gave to the English language, GloWbE does indeed show at a glance that such words as *brogues, galore, gob, slew, slogan*, *whiskey* are known and used throughout the world—both in terms of raw occurrences and of relativised or normalized frequencies per one million words (pmw), as shown in **Figure 11** for *galore*:

The worldwide spread of *galore* shown by of the GloWbE distribution confirms the DHE claim that "this is one of the few words of Irish origin that have made their way into SE usage" (p. 109) particularly as a predicate adjective. In fact, the Irish frequency of 1.46 pmw is below the word's global average of 1.51 pmw, with a highest scores of 2.79 pmw in Malaysia and 2.33 in Jamaica. Here is an example from Northern Ireland:

He has got cups **galore**, that man up there (NITCS Text 20, Doagh Co. Antrim) (cups referring to trophies won as prizes).[15]

Of all the words which Ireland has given the world, *brogue/brogues* must rank as a strong candidate. According to DHE, *brogue* (singular) refers to "the Irish way of speaking English" and, according to Murphy (1943), originated as an English word to describe an Irish accent, being first recorded in the early 16th century with reference to a parrot which could imitate accents including Irish English (Hickey, 2007, p. 7). Nowdays, its meaning has been generalized to refer to the whole way Irish people are perceived to speak English "unintelligibly (as a result of contamination from Irish syntax and vocabulary)" (DHE).[16]

In **Figure 12**, the GloWbE map shows how widespread *brogue* and *brogues* are, but not surprisingly that most occurrences remain in Ireland.

Another word claimed to have been given to the English language from Ireland is the noun *slew*, meaning "a large number or amount of something" < Irish *slua* "a crowd, a host, a multitude" (DHE) notwithstanding the OED: "originally U.S."

---

[12]The other four types are "Traditional L1 varieties", "indigenized L2 variety", "English-based pidgins", and "English-based creoles".

[13]Cf. for instance, those carried out at different times by Henry (1995) (for her 1995 book *Belfast English and Standard English*) or Ray Hickey (for his Survey of Irish English Usage, reported in Hickey, 2007).

[14]https://www.english-corpora.org/glowbe/ (last accessed January 05, 2022).

[15]The example is from the *Northern Ireland Transcribed Corpus of Speech* (Kirk, 2004).

[16]By contrast, *brogues* (plural) refers to a pair of strong leather shoes with a pattern of holes and stitches, supposedly from Irish *bróg* "a shoe", or else narrowed from Old Norse *brók* ("trousers", whence *breeches*) (cf. Bergin, 1943, cited by Hickey, 2007, p. 7), but it is unlikely that the reference to shoes has transferred to speech to suggest that the way the Irish speak English was "as if they had a shoe on their tongue" (DHE) (clearly a false etymology!). It is only this shoe meaning which features in the entry in *A History of Ireland in 100 words* (Arbuthnott et al., 2019, p. 86–87).
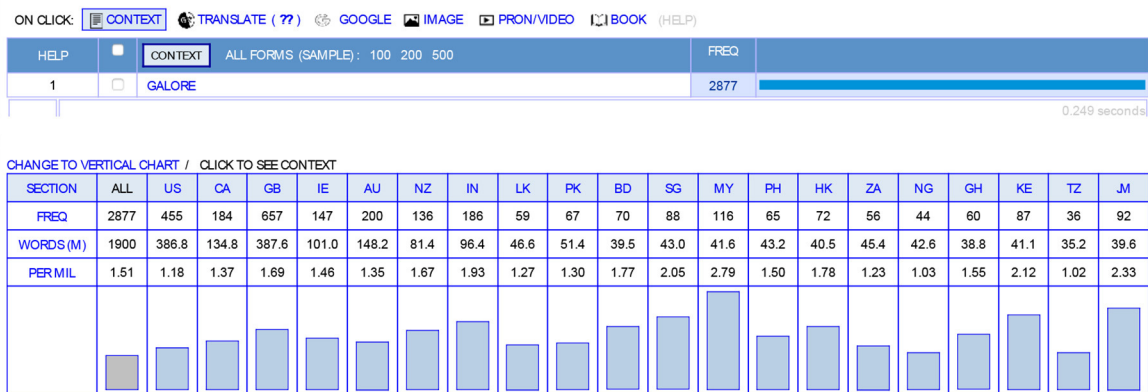
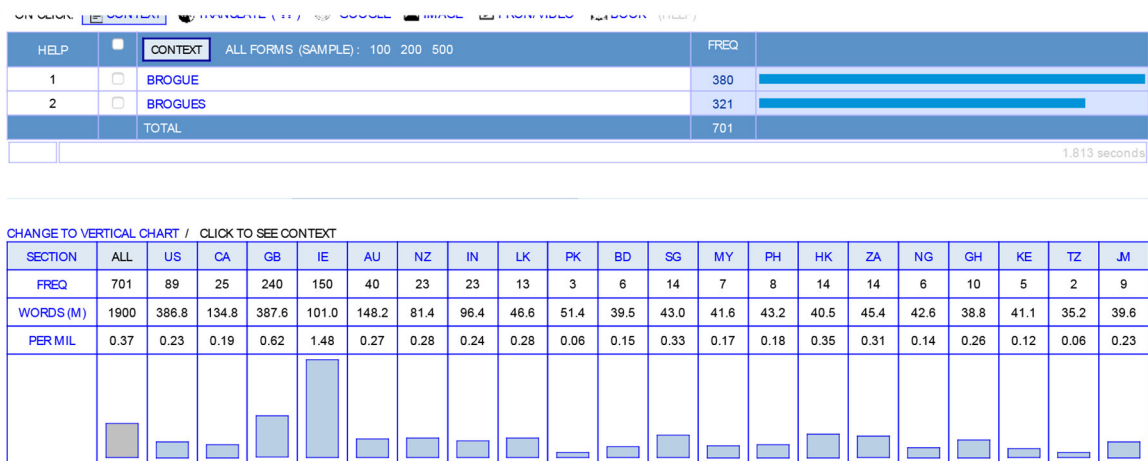**FIGURE 11 |** Distribution of *galore* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

FIGURE 12 | Distribution of *brogues* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

In **Figure 13**, the GloWbE map shows a widespread distribution, with Ireland (with a pmw of 1.40) far from predominating compared with Jamaica (pmw 6.44), India (pmw 3.53) and the USA (pmw 3.39).

The word *slogan* in English is claimed as often to come from Scottish Gaelic as it is from Irish. Originally it meant "a war cry", from the Irish *slua-ghairm* or Scottish Gaelic Gaelic *sluagh-ghairm* "a battle cry". However, it is its metaphoric sense of "a distinctive word or phrase used by a political or other group", first attested 1,704, according to Etymonline, which has gone round the world, as the map in **Figure 14** shows.

Slogan would appear now to be most frequent in Ghana (pmw 12.67) and Malaysia (pmw 10.76), compared with a pmw of 5.12 in Ireland.

The final word for consideration as an Irish contribution to English is *whiskey/whiskeys*, from Irish *uisce* "water" + *beatha* "life", whence *eau de vie*, itself probably a loan-translation of Medieval Latin *aqua vitae*. In Scotland, the word is spelled *whisky*, which came through Scottish Gaelic from Irish as well.

In **Figure 15**, the GloWbE map shows not surprisingly that, although nowadays an international beverage, *whiskey* is most common in Ireland itself (pmw 17.68).

As for the Irishisms that are favorites and much-loved at home, GloWbE can reinforce such local preferences by showing (as in **Figure 16**) that a word such as *langers*, which refers to the male penis and by extension to "a stupid or contemptible person" or to "someone who is drunk", and which is regarded as a slang term of contempt and abuse, occurs almost exclusively in Ireland and is non-existent in many countries.
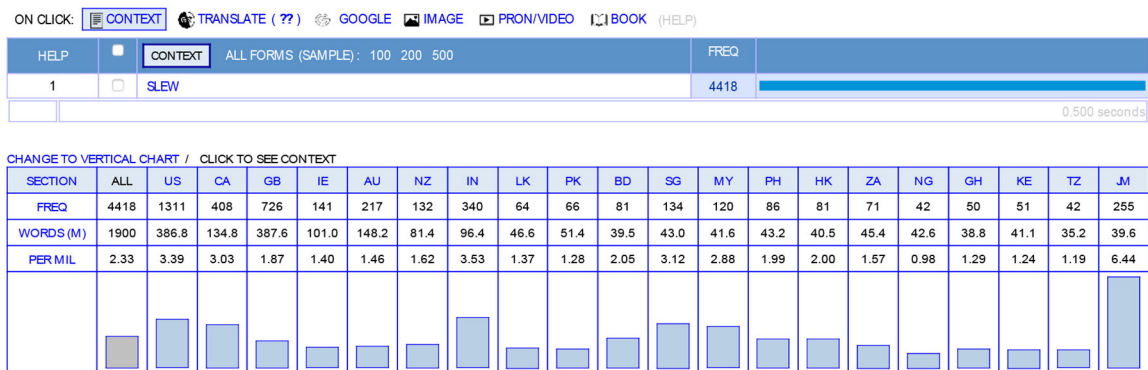
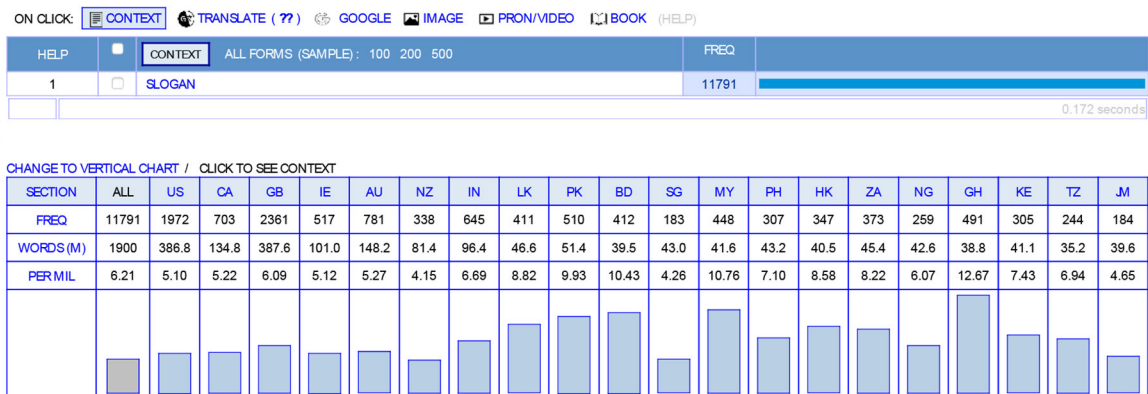**FIGURE 13** | Distribution of *slew* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

FIGURE 14 | Distribution of *slogan* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

A further favorite is *Leprechaun*. Few folk motifs are more indexical of the Irish than *leprechauns*. According to DHE, a leprechaun is "a dwarflike sprite; an industrious fairy seen at dusk or in the moonlight mending a shoe".[17] For *Slanguage*, *leprachauns* are "characteristic of perceived Oirish [sic] qualities" [i.e., "excessive or exaggerated display of perceived national characteristics", and its use is generally derogatory]. In **Figure 17**, the GloWbE map shows that the country of most occurrence is indeed Ireland—not unexpectedly—at pmw 2.84, but that, with the exception of Tanzania, there are occurrences of *leprechaun(s)* in every other of the 20 countries surveyed. After Ireland, *leprechaun* occurs most frequently in, perhaps surprisingly, Singapore (pmw 0.63), followed by the USA (pmw 0.48) and then

GB (pmw 0.40). The word may be Irish in origin[18] but the display shows that *leprechaun(s)* has come to be used throughout the world, albeit in small numbers.

A similar distribution occurs for another indexical word of the Irish: *eejit/eejits*, referring to "a silly person", "a fool", "an idiot", an Irish (especially Dublin) spelling pronunciation of *idiot* but, according to DHE, "less pejorative than StE *idiot*".

---

[17]But cf. Arbuthnott et al.. (2019, p. 117–118) for an alternative view that leprechauns are not Irish at all.

[18]According to DHE *leprechaun*, is from Irish *leipreachán*, and accordingx to etymonline.com, *leprechaun*, is from Irish *lupracan*, a metathesis of Old Irish *luchorpan*, which traditionally is explained as literally "a very small body," from *lu*"little, small" (from PIE root *legwh-*"not heavy, having little weight") + *corpan*, diminutive of *corp*"body," from Latin *corpus*"body" (from PIE root *kwrep-*"body, form, appearance"). The variant *leithbragan* is probably an Irish folk etymology, from *leith*"half" + *brog* "brogue" because the spirit was supposed to be always employed in making or mending a single shoe – whence the common image.

FIGURE 15 | Distribution of *whiskey/whiskeys* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

In **Figure 18**, the GloWbE map confirms the word *eejit(s)* as an Irishism (pmw 2.69), with few occurrences elsewhere except for GB (pmw 0.30), where Irish influence is strongly felt.

A further favorite word is *donnybrook(s)*. According to DHE, *donnybrook* is "a riotous assembly, a free-for-all" deriving from the place-name *Donnybrook*, "a village near Dublin, now a respectable suburb, site of an uproarious week-long fair dating from at least the 13th century, discontinued in the middle of the 19th century". In **Figure 19**, the GloWbE map shows that *donnybrook(s)* hardly occurs outside of Ireland at all—hardly surprising, perhaps, because of its cultural associations—and thus confirming it as another Irishism.

The word *Tory* (plural *Tories*), originally meaning "a bandit" or "outlaw", "an Irish vagabond robber or rapparee" (according to *Slanguage*), came to acquire political connotations to refer to "an advocate for absolute monarchy and church power". Despite its Irish origins, the concept of a *Tory* has nowadays narrowed and relates predominantly to members of the Conservative Party in Great Britain or in Canada. Little surprise that this word has become a culturalism which is shown, in **Figure 20**, on the GloWbE map to predominate in Canada (pmw 16.61) and especially in GB (pmw 77.94) but is relatively scarce elsewhere, including Ireland (pmw 7.67).

The final candidate for a favorite Irish word is *boycott(s)*, meaning "refusal to cooperate", "to exclude from all social or commercial intercourse".[19]

In **Figure 21**, the GloWbE map shows that *boycott* occurs throughout the world, most frequently in Sri Lanka (pmw 13.70) and Pakistan (pmw 11.27). It has also been borrowed into German as *der Boykott*, for instance. Despite its source, the word can hardly be claimed to be an Irishism any more, something which could not have been predicted without such comparable corpus resources.

Thus the claims about language based on common knowledge and regularly made by journalists such as Mullally and Phelan may now be verified by using resources such as GloWbE. Using GloWbE's extensive, suitably marked-up data as well as its instantaneous search and display capabilities, we have been able to show how some words are confirmed—visually confirmed on the display maps—as Irishisms whereas others which have originated in Ireland are nowadays to be found throughout the English-speaking world.

STATISTICAL ANALYSES OF IRISH ENGLISH AS A WORLD ENGLISH

As we have seen above with the *Mouton Atlas* and eWAVE, attempts using morpho-syntactic data or lexical data to relate world Englishes to each other have proceeded on an item by item basis, whether the *after*-perfect, verb complementation patterns or lexical phenomena. Whereas there has been a focus on national varieties, when ICE corpora have been used for comparison, attention is usually paid to mode differences of distribution or function as between speech and writing, and occasionally to the various spoken discourse situations or written registers in the mode components. Registers

---

[19]As an eponym, *boycott* was made famous after a certain Capt. Charles C. *Boycott* (1832–1897), a landowner in County Mayo, who refused to lower rents for his tenant farmers, so that his "oppressed tenants and workmen as well as local business people embarked on a celebrated policy of non-cooperation on his estate in the autumn of 1880" according to DHE.

FIGURE 16 | Distribution of *langers* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

FIGURE 17 | Distribution of *leprechaun* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

FIGURE 18 | Distribution of *eejit/eejits* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).
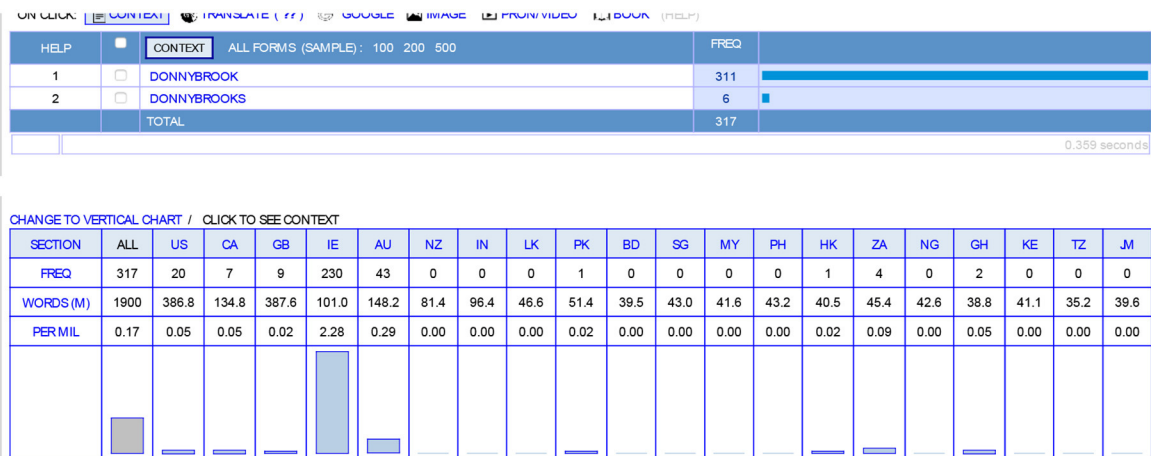
FIGURE 19 | Distribution of *donnybrook* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

FIGURE 20 | Distribution of *Tory/Tories* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).

have been variously identified on external social grounds such as the constellation of speakers or their function in communicating in the discourse situation, or whether the written texts served informational, instructional, persuasive or creative functions. Against this background, and with the availability of computational resources, there has emerged an approach of register comparison based on groupings of internal features and their implicit linguistic functions. The approach came to be known as "multi-dimensional factor analysis", reflecting that groups of features can share the same linguistic function or "dimension", that there are a number of key linguistic dimensions, and that consequently

each dimension becomes realized through the presence of a goodly number of occurrences of each feature or, in some cases, an absence of certain other features. The method came to prominence by Biber (1988) ground-breaking study of 1988; now, after revisions by Biber himself, there is the further ground-breaking study by Bohmann (2020). Bohmann's approach has culminated in 10 dimensions (such as "involved vs. informational production" or "collaborative communicative orientation" or "conceptual vs. concrete informational focus" and so on) (p. 94–99) each indexical with a selection from some 236 linguistic forms including morpho-syntactic forms and affixes of word-formation ("the factors") (p. 73–81). These
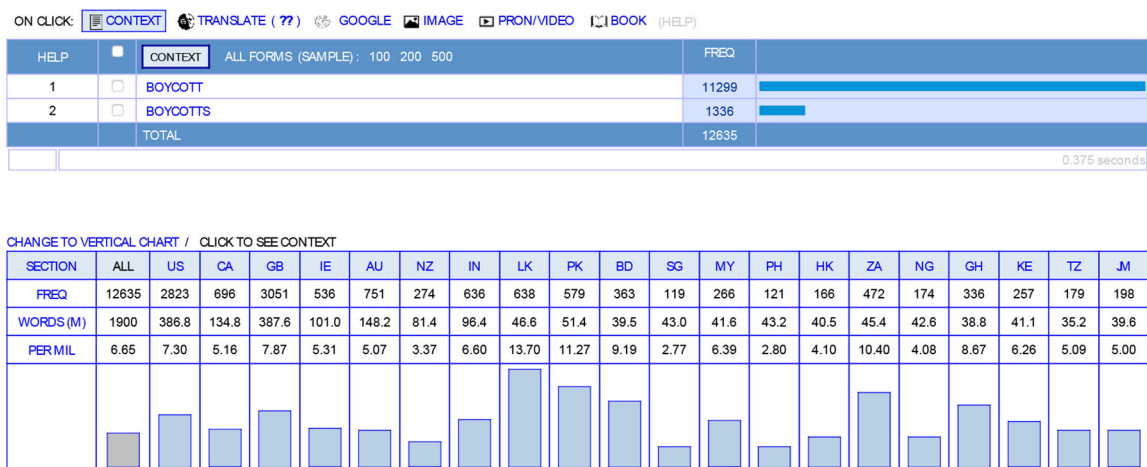
FIGURE 21 | Distribution of *boycott(s)* in the GloWbE Corpus (from https://www.english-corpora.org/glowbe/).
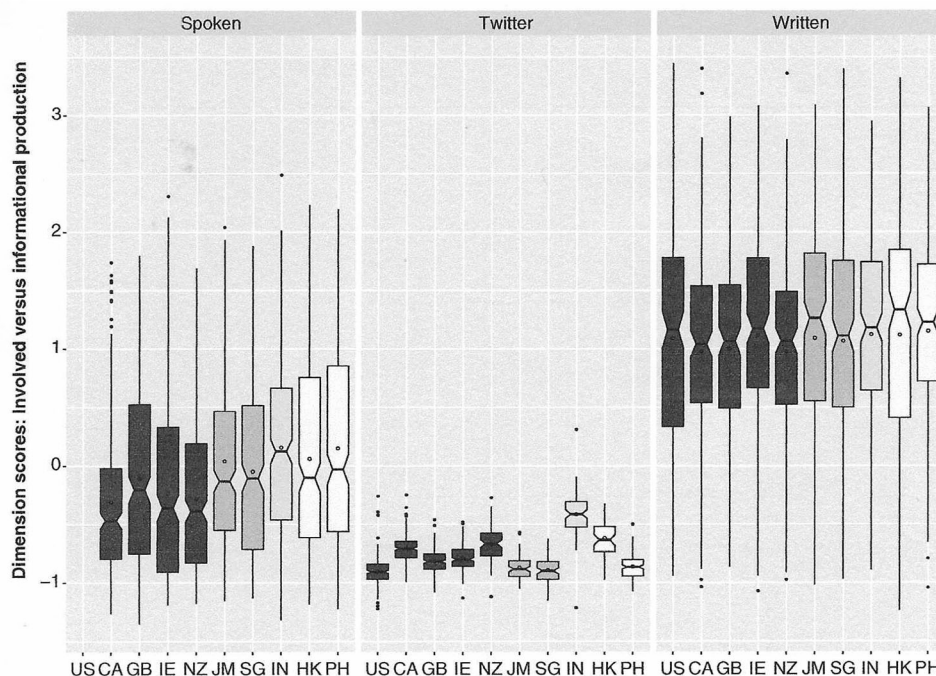

FIGURE 22 | Bohmann's "Distribution of dimension scores across varieties and modalities; INVOLVED VS. INFORMATIONAL PRODUCTION" (from Bohmann, 2020, p. 109).

factors have been annotated by Bohmann in respect of their respective dimensions in some ten ICE-corpora, including ICE-Ireland. Using multi-dimensional factor analysis, with regard to the distribution of scores for the dimension of "involved vs. informational production", Bohmann observes that the scores reveal differences of style and comments that "distinctly low average values mean[...] a very clear interpersonal and structurally simple style, [which] can be noted for Canadian

English, Irish English and New Zealand English, whereas the highest values are found for Indian English, Philippine English, and to a lesser extent Jamaican English and Hong Kong English." (p. 109) Bohmann continues: "This grouping [which Bohmann depicts in his Figure 5.3, p. 109, reproduced here as **Figure 22**] suggests a slight distinction between Inner and Outer Circle varieties in terms of the "involved vs. informational production" such that the former tends to be more

interpersonal and structurally simpler than the latter in oral communication" (p. 109).

In such ways, Bohmann is able to demonstrate connections between the relationships within the Kachru model, on the one hand, and the evidence of variation between corpora on the basis of the distribution of a large amount of features on the other.

In a further investigation, ICE-corpora together with "a custom-made corpus of geolocated Twitter messages from around the English-speaking world" collected between 2011 and 2015 (known as TwICE) (Bohmann, 2020, p. 65) are compared in terms of their stage of development in the Schneider Dynamic Model. All components are at least at stage 3, with Ireland, we recall, at stage 5 ("differentiation"). Overall, Bohmann concludes that multi-dimensional factor analysis of the total data set for each country can be aligned with the theoretical insights of the Dynamic Model's evolutionary stages (p. 192). In his Figure 6.4 (p. 131), reproduced here as **Figure 23**, he shows that Irish English clusters with the other L1 varieties, thereby justifying its categorization as a stage-5 differentiated and L1 variety.

Last but not least, to identify characteristics which varieties of English have in common, let us now to turn to the most abstract level of theoretical syntactic argumentation, linguistic universals.[20] A typology of seven universals is presented by Szmrecsanyi and Röthlisberger (2020, p. 542):

- GENUINE UNIVERSALS (e.g., "all languages have vowels"),
- TYPOVERSALS ("features common to languages of a specific typological type"),
- PHYLOVERSALS ("features shared by a family of genetically-related languages"),
- AEROVERSALS ("features common to languages which are in geographical proximity to each other"),
- VERNACULAR UNIVERSALS ("features that are common to spoken vernaculars"),
- ANGLOVERSALS ("features that recur in vernacular varieties" of, in this case, English),
- VARIOVERSALS ("features recurrent in language varieties with a similar socio-history, historical depth, and mode of acquisition").

Based on the author's knowledge and observation, features of Irish English may be accounted for in several of these typologies. For starters, universally, it has vowels. Typoversally, it is basically a SVO language, although the medial-object perfect construction is an SOV construction (cf. Filppula, 1999; Hickey, 2007). Phyloversally, the extended temporal reference in uses of the past tense and present tense relates it to other Germanic languages (cf. Siemund, 2004). Aeroversally, some features which have transferred from Irish are found in Highland English, having transferred from Scottish Gaelic (cf. Sabban, 1981). Vernacularly, Irish English shares the features postulated by Jack Chambers (2004) such as double/multiple negation (usually used emphatically), angloversally (e.g., unmarked use of adjectives as adverbs) (cf. Szmrecsanyi and Kortmann, 2009),

---

[20]This isn't to be confused with the notion of universals in language acquisition, as advanced by Filppula (1999), Siemund (2004), and Filppula et al. (2009), among others, as an explanation for the distinctiveness of Irish English morpho-syntax.
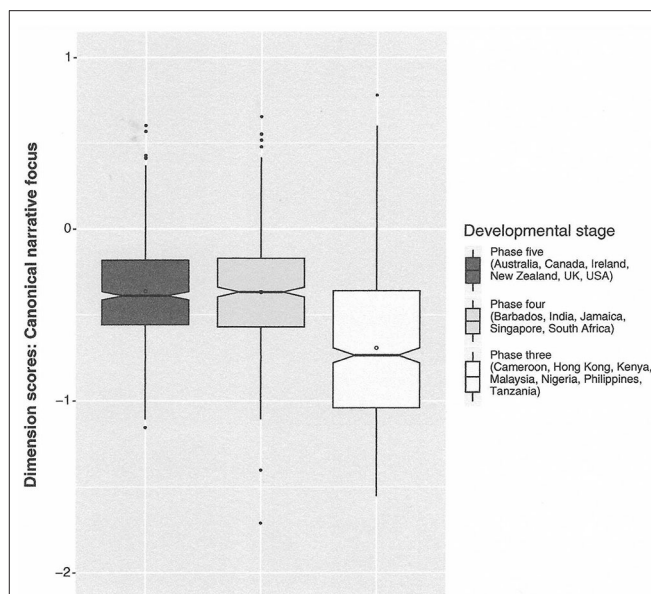

FIGURE 23 | Bohmann's "Distributions of dimension scores for countries in different phases according to the Dynamic Model: ICE data only: CANONICAL NARRATIVE FOCUS" (from Bohmann, 2020, p. 131).
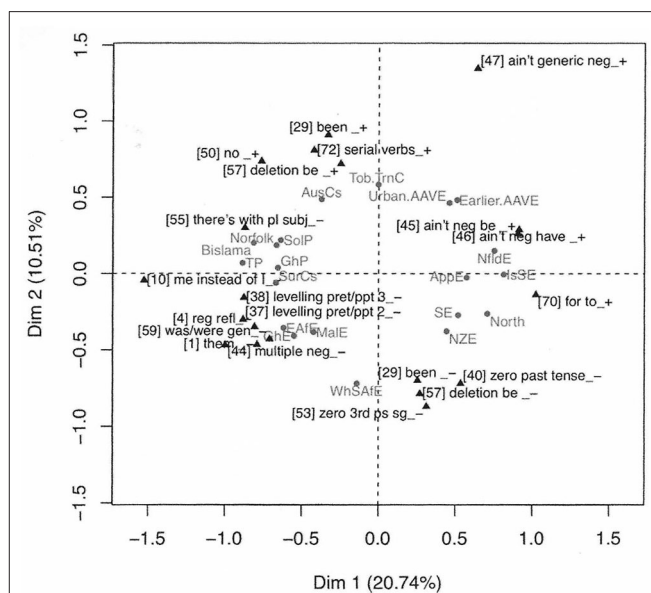
FIGURE 24 | Szmrecsanyi and Röthlisberger's multiple correspondence analysis for world Englishes (from Szmrecsanyi and Röthlisberger, 2020, p. 545).

and varioversally (as in the leveling of past tense and past participle verb forms or object pronouns used as subjects) (cf. *Mouton Atlas*/eWAVE).

Using the data that was published in the earlier version of the *Mouton Atlas* and eWAVE, namely the *Handbook of Varieties of English* (HVE) (Kortmann et al., 2004), Szmrecsanyi and his collaborators reach similar conclusions to Bohmann.

Szmrecsanyi and Röthlisberger (2020) present a multiple correspondence analysis for world Englishes. The dialectometric technique used by them is a multiple correspondence analysis, which explores how features are associated with each other in order to establish the extent to which the varieties relate to each other. Thus, had it been plotted in their Figure 23.3 (p. 545), reproduced here as **Figure 24**, Irish English would almost certainly have appeared in the same lower right-hand quadrant of the plot as dialects of British English with features such as the *for-to* infinitive and absences of deleted *be*. As Szmrecsanyi and Röthlisberger (2020, p. 545) comment: "the most important dimension of variation (Dim 1) pits native varieties (right) against pidgin/creoles and L2 varieties (left). The vertical dimension (Dim 2) appears to be capturing a language-externally defined contrast between orientation toward North American English (top) vs. orientation toward British English (bottom)."

A further concern of Szmrecsanyi and Röthlisberger (2020) is a variety's degree of analyticity and syntheticity, which they measure as an index of free vs. bound grammatical markers per word, across a geographically widespread range of varieties of English. It appears that L2 varieties have a greater frequency of analyticity markers whereas traditional L1 varieties such as Irish English have a greater degree of syntheticity. A plot of analyticity/syntheticity (their Figure 23.4, p. 548, a differently calibrated version of Siegel et al. (2014): Figure 2, reproduced here as **Figure 25**) shows Irish English between Southwest of England dialects and Scottish Lowlands Dialects and surrounded by L1 countries—all set on the plot quite apart from the likes of Tok Pisin, Hawai"i Creole, Hong Kong English and Singaporean English.

A further typological dimension of comparison between varieties of world Englishes is that of complexity vs. simplicity. For many linguists all languages are equally complex. Nevertheless, some have argued that creoles are much less complex grammatically than their lexifier languages. Thus, it would appear that complexity is variable after all. The question thus arises of how complex Irish English is. Within varieties of English, it has been shown that L2 varieties are simpler—not least because of imperfect learning or the simplification of rules, or the regularization of constraints. In a nutshell, "rule simplicity predicts L2 simplicity and vice versa" (p. 551). In Szmrecsanyi and Röthlisberger (2020, p. 551), Figure 23.5, reproduced here as **Figure 26**, combines x-axis rule simplicity with y-axis the number of L2 simple features (what they call "acquisition simplicity"). On the basis of 76 morpho-syntactic features known to simplify in some varieties plotted against features "known to recur in interlanguage varieties" (p. 551), two points emerge in the plot: "rule simplicity predicts L2 simplicity"; and that with fewer simplyfing and simple features, L1 varieties cohere in the middle—and thus Irish English (IrE) alongside northern English dialects (North) and colloquial American English (CollAmE)—and are thus set apart from pidgins and creoles on the top periphery, with all their simplifications.

The provision of lists of features as in the Kortmann initiatives (HVE; *Mouton Atlas* and eWAVE) as well as the advent of the ICE corpora have enabled previously unimagined investigations
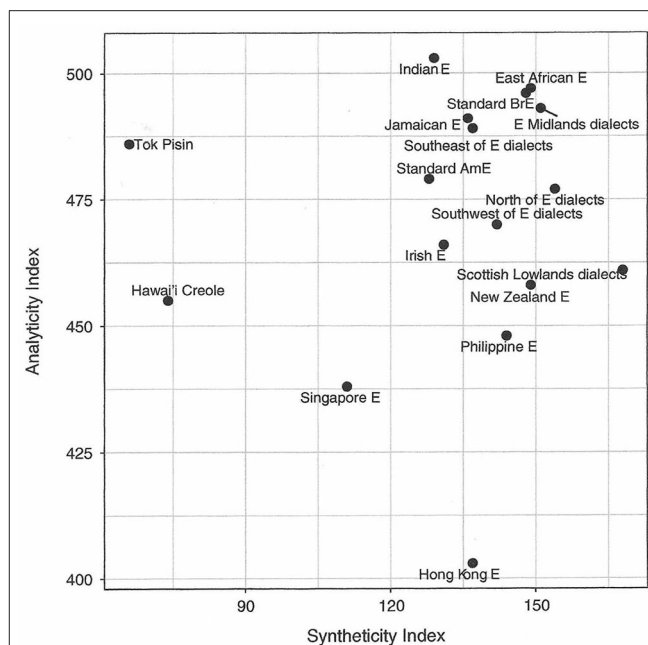

FIGURE 25 | Szmrecsanyi and Röthlisberger's plot for syntheticity and analyticity among world Englishes (from Szmrecsanyi and Röthlisberger, 2020, p. 548; after Siegel et al., 2014; **Figure 2**).
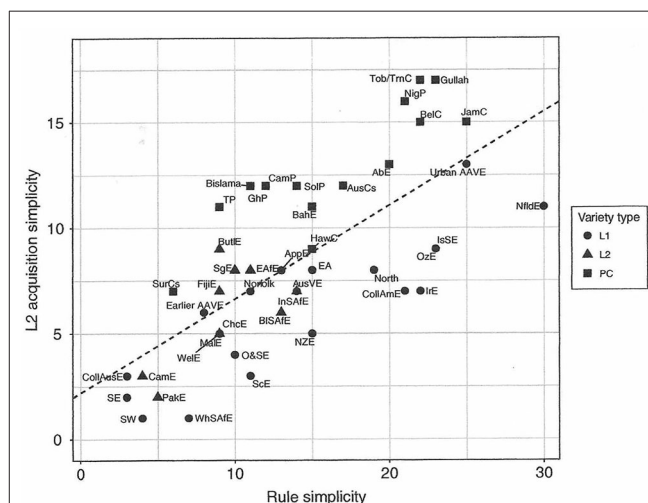
FIGURE 26 | Szmrecsanyi and Röthlisberger's simple vs. simplifying varieties among for world Englishes (from Szmrecsanyi and Röthlisberger, 2020, p. 551).

into the nature of and relationship between national varieties of world Englishes. The approaches using datasets have largely been computer-based, whether for qualitative analyses and interpretations, or for quantitative comparisons. Irish English has been at the center of all these enquiries. What emerges recurrently is confirmation of the status of Irish English as an L1 variety, however exotic the Celticity of its background might appear (cf. Kirk and Kallen, 2006). An L1 variety is, of

course, a general categorical labeling which doesn't exclude the possibility of koinés (Onysko, 2016) from a contact-typological point of view, as L1 Englishes undergo further dialect contact, as shown above. Nevertheless, on the basis of increasingly global statistical analyses of the totality of the standardized data available in the ICE-Ireland corpus, the recurrence of L1 confirmation may render redundant some of the above top-down classifications of Irish English such as "sub-regional semi-standard" (Görlach, 1990), "new" (Mesthrie and Bhatt, 2008), "central" (Mair, 2013), "high contact" (eWAVE), and "differentiated" (Schneider, 2007), and "postcolonial" (Schneider, 2007). Against the results generated by multi-dimensional factor analysis or multiple correspondence analysis, such externally-motivated claims about the categorical or paradigmatic status of a national variety are difficult to substantiate on the basis of a corpus-based study.

CONCLUDING REMARKS

This chapter, which explores the relationship between Irish English and world English(es), has raised many issues. It has raised the need to clarify the notion of "Irish English" recurrently, and to distinguish between the vernacular and the standardized variety, and also the need for a critical need for awareness between the two as well as for points along any postulated continuum between the two. The chapter contends that it is because of the presence of both Hiberno-English features as well as features from Scots that the wider reality of educated public discourse in Ireland is not fully standardized. It further contends that, when the status of the data is hypothesized, the substantiality of several possibilities regarding status not only within Ireland and between Ireland and Great Britain, but also between Irish and varieties of world Englishes anywhere becomes possible.

The chapter onsiders the status which has been accredited to Irish English in a number of top-down models of varieties of English around the world, by McArthur (1987), Görlach (1990), Mesthrie and Bhatt (2008), Mair (2013), Siemund (2013), and Onysko (2016), among a few others, on the one hand, and on the other by Schneider (2007) Dynamic Model. The general challenge for all models of world Englishes is their demonstration and vindication of hypotheses on the basis of empirical corpus data. Whereas there have been few attempts hitherto, the review of them by Hundt (2020) plausibly concludes that the gap would appear unbridgeable. As the present article goes on to show, there is, however, no shortage of data available—ranging from the ever-growing suite of ICE-corpora to the eWAVE database and the GloWbE corpus as resources of both non-standardized as well as standardized features, for comparison, and for statistical evaluation. A further section deals with the status of lexical items originating in Ireland and, by investigating the GloWbE corpus, is able to discriminate between those which have remained there as Irishisms and those which have spread their wings and are now to be found in almost every variety around the world. These global comparisons of individual items lead into the final part of the chapter, which is concerned with various statistical attempts to assess and relate Irish English data to a great range of varieties

worldwide, including pidgins and creoles. Whether ICE-corpora are used for data or the list of observational data behind eWAVE, or whether multi-dimensional factor analysis is used on the standardized data, or whether multiple correspondence analysis is used on the reported observations of Hiberno-English data, Irish English becomes repeatedly vindicated as an L1 or (in the Schneider, 2007 model) a stage-5 differentiated, variety, *tout court*.

What this study also shows, in the context of national varieties of Englishes, is the difficulty of reconciling top-down, theoretical, deductive approaches with bottom-up, empirical, inductive approaches, and the need for a critical awareness and for qualification in each approach if the two are ever to meet or be reconciled. Top-down can be neat and tidy, as models of trees and circles show, but real (and especially spoken) data are messy and often unamenable to tidy categorization. It is oversimplistic to think that one unique box or circle or hypothetical category in a schema can be superimposed on a national corpus or an exclusive national dataset. For all is confluence and flux. Irish English shares many lexical and morpho-syntactic features with other varieties of world Englishes, as shown, just as it also retains many other unique lexical and morpho-syntactic features, as also shown. Some of the latter may also be relatively rare. It is seldom a case of either/or, presence or absence, if ever. The reality of patterns of features and in varying frequencies is only shown by the approaches which have become developed within corpus statistics. From the plots and graphs which emerge as the output, as exemplified above, what comes to be shown are clusters of similarities whether of features which are indicative of varieties, as Bohmann shows (**Figures 22**, **23**), or of varieties which are indicative of the features underlying the statistics, as Szmrecsanyi and Röthlisberger show (**Figures 24**–**26**).

Nevertheless, we should not overly rely on corpus statistics as they are greatly abstracted from the original data and contexts, The popular opinion displayed in journalism can be can be corroborated by corpora—GloWbE distinguishes in lexis the global from the local or, in the case of minimal territorial expansion, the glocal, as does eWAVE for morpho-syntax. How really Irish is an Irishism can now be given a distributional answer beyond any local lexicographical treatment. National nuggets in international ore can indeed be identified and their co-existence discriminated by the research resources and tools now available. Irish English is a world English (and, despite everything, an L1 at that), but it also remains itself. Erin go bragh!

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

Kortmann for permission to reproduce results from eWAVE, and to Mark Davies for results from GloWbE. I am greatly indebted for very constructive feedback from two reviewers as well as the editors. A shorter version of the material used in this paper, with few reproductions, has appeared in the *Festschrift for Jeffrey L. Kallen* as Kirk (2021).

REFERENCES

Anderwald, L. (2002). *Negation in Non-Standard British English, Gaps, Regularizations, Asymmetries*. London: Routledge.

Arbuthnott, S., Mhaonaigh, M., and Toner, G. (2019). *A History of Ireland in 100 Words*. Dublin: Royal Irish Academy.

Barron, A., and Schneider, K. P. (2008). *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages*. Amsterdam: John Benjamins.

Bergin, O. (1943). *Bróg* "shoe". *Éigse*. 3, 237–239.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621024

Bohmann, A. (2020). *Variation in English Worldwide: Registers and Global Varieties*. Cambridge: Cambridge University Press. doi: 10.1017/9781108751339

Buschfeld, S., and Kautsch, A. (2017). Towards an integrated approach to postcolonial and non-postcolonial englishes. *World Englishes*. 36, 104–126. doi: 10.1111/weng.12203

Buschfeld, S., and Kautsch, A. (2020). *Modelling Worlds Englishes: A Joint Approach to Postcolonial and Non-Postcolonial Varieties*. Edinburgh: Edinburgh University Press. doi: 10.3366/edinburgh/9781474445863.001. 0001

Chambers, J. (2004). "Dynamic Typology and Vernacular Universals", in *Dialectology Meets Typology*, ed Kortmann, B. (Berlin: Mouton de Gruyter), p. 127–145.

Corrigan, K. P. (2020). *Linguistic Communities and Migratory Processes: Newcomers Acquiring Sociolinguistic Variation in Northern Ireland*. Berlin: de Gruyter. doi: 10.1515/9783110614190

Crystal, D. (1995). *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

Davies, M., and Fuchs, R. (2015). Expanding horizons in the study of world englishes with the 1.9 billion word global web-based english corpus GloWbE. *English World-Wide*. 36, 1–28. doi: 10.1075/eww.36.1.01dav

De Costa, P. I., Crowther, D., and Maloney J. (2019). Investigating World Englishes, Research Methodology and Practical Applications. *Abingdon: Routledge*. doi: 10.4324/9781315299716

Dehors, S. C.. (2018). Modelling World Englishes: Assessing the Interplay of Emancipation and Globalization of ESL Varieties. *Amsterdam: John Benjamins*. doi: 10.1075/veaw.g61

Diskin, C., and Levey, S. (2019). Going global and sounding local: quotative variation and change in L1 and L2 speakers of Irish (Dublin) English. *English World-Wide*. 40, 53–78. doi: 10.1075/eww.00022.dis

Dolan, T. P. (2014). *A Dictionary of Hiberno-English: The Irish Use of English*. Third Edition. Dublin: Gill and Macmillan.

Edwards, A. (2016). *English in the Netherlands: Forms, Functions and Attitudes*. Amsterdam: John Benjamins. doi: 10.1075/veaw.g56

Filppula, M. (1999). *A Grammar of Irish English: Language in Hibernian Style*. London: Routledge. doi: 10.2307/29742738

Filppula, M., Klemola, J., and Paulasto, H. (2009). *Vernacular Universals and Language Contacts: Evidence From Varieties of English and Beyond*. Abingdon: Routledge. doi: 10.4324/9780203883426

Filppula, M., Klemola, J., and Sharma, D. (2017). *The Oxford Handbook of World Englishes*. Oxford: Oxford University Press,

Görlach, M. (1990). *Studies in the History of the English Language*. Heidelberg: Carl Winter Universitätsverlag.

Görlach, M. (ed.). (1995). "Irish English and Irish Culture in Dictionaries of English", in *More Englishes, New Studies in Varieties of English (1988–1994)*. (Amsterdam: John Benjamins), p. 164–191. doi: 10.1075/veaw.g13.07iri

Greenbaum, S. (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gries, S., and Mukherjee, J. (2010). Lexical gravity across varieties of english. *Int. J. Corpus Linguist.* 15, 520–548. doi: 10.1075/ijcl.15.4.04gri

Haswell, C. (2013). A global model of english, how new modeling can improve the appreciation of english usage in the asia pacific region. *Asia Pacific World*. 4, 122–137. doi: 10.3167/apw.2013.040208

Henry, A. (1995). *Belfast English and Standard English: Dialect Variation and Parametric Setting*. Oxford: Oxford University Press.

Hickey, R. (2005). *Dublin English: Evolution and Change*. Amsterdam: John Benjamins.

Hickey, R. (2007). *Irish English: History and Present-day Forms*. Cambridge: Cambridge University Press.

Hundt, M. (2020). "Corpus-based approaches to world englishes," in *The Cambridge Handbook of World Englishes*, eds D. Schreier, M. Hundt, and E. W. Schneider (Cambridge: Cambridge University Press), 506–533.

Jenkins, J. (2014). *Global Englishes: A Resource Book for Students*. Abingdon: Routledge. [Third Edition of World Englishes].

Kachru, B. B. (1982). *The Other Tongue: English Across Cultures*. Urbana, Ill.: University of Illinois Press.

Kallen, J. L. (1996). "Entering Lexical Fields in Irish English", in *Speech Past and Present: Studies in English Dialectology in memory of Ossi Ihalainen*, eds Klemola, J., Kytö, M., and Rissanen, M. (Frankfurt am Main: Peter Lang), p. 101–129.

Kallen, J. L. (2013). *Irish English*, vol. 2: *The Republic of Ireland*. Berlin: de Gruyter.

Kallen, J. L. (2017). "English in Ireland", in *The History of English: Varieties of English*, eds Bergs, A., and Brinton, J. L. (Berlin: de Gruyter), p. 244–264.

Kallen, J. L., and Kirk, J. M. (2008). *ICE-Ireland: A User's Guide*. Belfast: Cló Ollscoil na Banriona.

Kirk, J. M. (2004). *Northern Ireland Transcribed Corpus of Speech*. Second Edition. Queen's University Belfast.

Kirk, J. M. (2021). "Irish English and World Englishes", in *Exploring the Landscapes of Irish English Research*, eds Lucek, S., and Amador-Moreno, C. (Abingdon: Routledge), p. 270–288. doi: 10.4324/9781003025 078-15

Kirk, J. M. (in press)."The Vocabulary of Irish English", in *Oxford Handbook of Irish English*, ed Hickey, R. (Oxford: Oxford University Press).

Kirk, J. M., and Kallen, J. L. (2006). "Irish Standard English: How Celticised? How Standardised?", in *The Celtic Englishes IV: The Interface between English and the Celtic Languages*, ed Tristram, H. L. C. Potsdam: (Potsdamer Universitätsverlag), p. 88–113.

Kirk, J. M., and Kallen, J. L. (2010). "How Scottish is Irish Standard English?" in *Northern Lights, Northern Words: Selected Papers from the FRLSU Conference, Kirkwall 2009,* Millar. Aberdeen: Forum for Research on the Languages of Scotland and Ireland. p. 178–213. Available online at: https://www.abdn.ac.uk/pfrlsu/volumes/vol2/

Kirk, J. M., and Kallen, J. L. (2011). "The Cultural Context of ICE-Ireland", in *Researching the Languages of Ireland*, ed Hickey, R. (Studia Celtica Upsaliensia: Uppsala University), p. 269–290.

Kirk, J. M., and Nelson, G. (2018). The international corpus of english project: a progress report. *World Englishes*. 37, 697–716. doi: 10.1111/weng.12350

Kirkpatrick, A. (2007). *World Englishes: Implications for International Communication and English Language Teaching*. Cambridge: Cambridge University Press.

Kirkpatrick, A. (2021). *The Routledge Handbook of World Englishes. Second Edition*. Abingdon: Routledge. doi: 10.4324/9781003128755

Klemola, J. (1994). Periphrastic DO in south-eastern dialects of British English. *Dialectologica et Geolinguistica* 2, 33–51 doi: 10.1515/dig.1994.1994.2.33

Klemola, J. (2002). "Periphrastic DO: Dialectal Distribution and Origins", in *The Celtic Roots of English, Studies in Language*, Vol. 37, eds. M. Filppula, J. Klemola, and H. Pitkänen (Joensuu: University of Joensuu, Faculty of Humanities), 199–210.

Kortmann, B., and Lunkenheimer, K. (2013). *The Mouton World Atlas of Variation in English*. Berlin: de Gruyter Mouton. doi: 10.1515/97831102 80128

Kortmann, B., and Lunkenheimer, K. and Ehret K. (2020). The Electronic World Atlas of Varieties of English. Zenodo. DOI, 10).5281/zenodo.3712132. [Available online at http://ewave-atlas.org, accessed May 09, 2022]

Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R. and Upton C. (2004). *A Handbook of Varieties of English*. Berlin: Mouton de Gruyter.

Lange, C., and Leuckert,. S. (2020). *Corpus Linguistics for World Englishes*. Abingdon: Routledge.

Ling Low, E., and Pakir A. (2018). *World Englishes: Rethinking Paradigms*. Abingdon: Routledge.

Macafee, C. I. (1996). *Concise Ulster Dictionary*. Oxford: Oxford University Press.

Mair, C. (2013). The world system of englishes: accounting for the importance of mobile and mediated vernaculars. *English Worldwide*. 34, 253–278. doi: 10.1075/eww.34.3.01mai

McArthur, T. (1987). The English languages? *English Today*. 3, 9–13. doi: 10.1017/S0266078400013511

Meierkord, C., and Schneider, E. (2021). *World Englishes at the Grassroots*. Edinburgh: Edinburgh University Press.

Melchers, G., Shaw, P., and Sundqvist, P. (2019). *World Englishes*. Abingdon: Routledge.

Mesthrie, R., and Bhatt, R. M. (2008). *World Englishes, The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

Migge, B. (2012). "Irish English and recent immigrants to Ireland", in *New Perspectives on Irish English*, eds Migge, B., and Ní Chiosáin, N., (Dublin: John Benjamins Publishing Company), p. 311–326. doi: 10.1075/veaw.g44.15mig

Modiano, M. (1999). International English in the global village. *English Today*. 15, 22–28. doi: 10.1017/S026607840001083X

Mukherjee, J., and Gries, S. (2009). Collostructional nativisation in new englishes, verb-construction associations in the international corpus of English. *English World-Wide*. 30, 27–51. doi: 10.1075/eww.30.1.03muk

Mullally, U. (2017). *Up to 90, Ireland in Our Favourite Words and Phrases*. Dublin: The Irish Times.

Murphy, G. (1943). English "brogue" meaning "Irish accent". *Éigse*. 3, 231–236.

Nestor, N., Ni Chasaide, C., and Regan, V. (2012). "*Discourse "like" and social identity, A case study of Poles in Ireland*", in New New Perspectives on Irish English, eds Migge, B., and Chiosáin, N. (Poland: John Benjamins Publishing company), p. 327–354. doi: 10.1075/veaw.g44.16nes

Onysko, A. (2016). Modelling world englishes from the perspective of language contact, *World Englishes*. 35, 196–220. doi: 10.1111/weng.12191

Onysko, A. (2021). *Research Developments in World Englishes*. London: Bloomsbury.

Phelan, K. (2016). *13 Words the Irish Gave the English Language*. Dublin: Irish Times.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Ronan, P. (2020). "English in Ireland, Intra-territorial Perspectives on Language Contact", in *Modelling World Englishes: A Joint Approach to Postcolonial and Non-Postcolonial Varieties*, eds Buschfeld, S., and Kautsch, A. (Edinburgh: Edinburgh University Press), p. 322–346. doi: 10.3366/edinburgh/9781474445863.003.0015

Sabban, A. (1981). *Gälisch-Englischer Sprachkontakt*. Heidelberg: Julius Groos Verlag.

Saraceni, M. (2021). *Bloomsbury World Englishes*, vol. 1: *Paradigms*, vol. 2: *Ideologies*, vol. 3: *Pedagogies*. London: Bloomsbury.

Schneider, E. (2003). The dynamics of new Englishes: from identity construction to dialect birth. *Language*. 79, 233–281. doi: 10.1353/lan.2003.0136

Schneider, E. (2004). How to trace structural nativization: particle verbs in world englishes. *World Englishes*. 23, 227–249. doi: 10.1111/j.0883-2919.2004.00348.x

Schneider, E. (2007). *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press.

Schneider, E. (2014). New reflections on the evolutionary dynamics of world englishes. *World Englishes*. 33, 9–32. doi: 10.1111/weng.12069

Schneider, E. (2020). *English Around the World: An Introduction. Second Edition*. Cambridge: Cambridge University Press.

Schreier, D., Hundt, M., and Schneider, E. (2020). *The Cambridge Handbook of World Englishes*. Cambridge: Cambridge University Press.

Seoane, E., and Suárez-Gómez, C. (2016). *World Englishes: New Theoretical and Methodological Considerations*. Amsterdam: John Benjamins.

Share, B. (2014). *Slanguage: A Dictionary of Slang and Colloquial English in Ireland. Third Edition*. Dublin: Gill and Macmillan.

Siegel, J., Szmrecsanyi, B., and Kortmann, B., (2014). Measuring analyticity and syntheticity in creoles. *J. Pidgin Creole Lang*. 29, 49–85. doi: 10.1075/jpcl.29.1.02sie

Siemund, P. (2004). "Substrate, superstrate and universal: perfect constructions in Irish English". in *Dialectology Meets Typology*, Kortmann, B. Berlin: Mouton de Gruyter. p. 401–434.

Siemund, P. (2013). *Varieties of English: A Typological Approach*. Cambridge: Cambridge University Press.

Siemund, P., Davydova, J., and Maier, G. (2012). *The Amazing World of Englishes: A Practical Introduction*. Berlin: de Gruyter.

Siemund, P., and Leimgruber, J. R. E. (2021). *Multilingual Global Cities: Singapore, Hong Kong, Dubai*. Abingdon: Routledge.

Strevens, P. (1980). *Teaching English as an International Language: From Practice to Principle*. Oxford: Pergamon Press.

Suárez-Gómez, C., and Seoane, E. (2015). *Englishes Today: Multiple Varieties, Multiple Perspectives*. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Szmrecsanyi, B., and Kortmann, B. (2009). "Vernacular universals and angloversals in a typological perspective", in *Vernacular Universals and Language Contacts: Evidence from Varieties of English and beyond*, eds, Filppula, M., Klemola, J., and Paulasto, H. (London: Routledge), p. 33–53.

Szmrecsanyi, B., and Röthlisberger, M. (2020). "World Englishes from the Perspective of Dialect Typology", in The Cambridge Handbook of World Englishes, eds Schreier, D., Hundt, M., and Schneider, E. W. (Cambridge: Cambridge University Press), p. 534–558. doi: 10.1017/9781108349406.023

Van Coetsem, F. (1988). *Loan Phonology and the two Transfer Types in Language Contact*. Berlin: de Gruyter. doi: 10.1515/9783110884869

Van Coetsem, F. (2000). *A General and Unified Theory of the Transmission Process in Language Contact*. Heidelberg: Carl Winter Universitätsverlag

Check for updates

Malay and English Language Contact in Social Media Texts in Brunei Darussalam and Malaysia

James McLellan*

Faculty of Arts and Social Sciences, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei

This paper focuses on aspects of language contact, specifically Malay and English in the domain of social media. Key components of the theoretical framework are world Englishes being by definition code-mixed varieties, and the complementary notions of nativization (of English) and englishization (of Malay, in this case). Texts examined and analyzed are from Whatsapp chats and groups, with the consent of all participants, and from public social media sites in the Malay world, mostly Negara Brunei Darussalam (henceforth Brunei) but also Malaysia. Brunei is known for having a very high percentage of social media use per head of population, with especially high levels of use of Instagram and Facebook, as well as discussion forums such as Reddit. In their social media interactions Bruneians and Malaysians have a range of language choices, from monolingual English to monolingual Malay, and varying degrees of code-mixing or translanguaging. Many Bruneians and Malaysians are multilingual, and thus may have more than two languages as resources to draw on. Analysis of threads of discussion forum postings on the same topic demonstrate the multilingual repertoire of participants, for whom any of the available language choices are unmarked. This is in part owing to the use of English as one medium of education alongside Malay: consistently in Brunei since 1985, inconsistently in Malaysia since 1963. The conclusion of the paper raises two questions: whether it is valid to posit the language of social media as a new variety comprising both local and global influences and inputs, and whether social media is a driver of change in varieties of English in Southeast Asia.

Keywords: contact, Malay-English, nativization, englishization, social media, translanguaging

INTRODUCTION

Bearing in mind the title and the importance of this research topic, "Englishes in a Globalized World: Exploring Contact Effects on Other Languages," it is necessary at the outset to state that this article is not just about Englishes. It attempts to highlight the salience of the "other languages." World Englishes are by definition code-mixed or translanguaging varieties (McLellan, 2020), and this article draws on the *kueh lapis* (Malay, "layer cake") analogy used by Haji-Othman and McLellan (2014) with reference to Brunei. This was developed as a means of showing that English, in Brunei and Malaysia, as elsewhere in multilingual societies, is just one language among many, akin to the many layers of a layer cake, a delicacy in Borneo. Haji-Othman (2012, p. 175–190) sums up the issue succinctly in the Brunei context in his chapter entitled "Is it always English? Dueling aunties in Brunei Darussalam," which aptly envisages English and Malay as

two dueling aunties competing for influence. The complementary notions of nativization (of English) and englishization (of local languages), developed by Kachru (2005, p. 113–117) are adopted as a basis for discussion of features of Malay/English language contact phenomena. The interconnected notions of code switching, code mixing, language alternation and translanguaging are outlined below.

The Context: Brunei

Negara Brunei Darussalam (henceforth Brunei) is a small Malay Islamic Sultanate located on the north-western coast of the island of Borneo, with a coastline of about 160 km on the South China Sea. It is surrounded on the other three sides by the East Malaysian state of Sarawak, which also divides the Temburong District from the other three Brunei administrative districts, Brunei-Muara, Tutong and Belait. The total land area is 5,675 km^2. The population of about 453,600 (http://www.deps.gov.bn/SitePages/Population.aspx) is concentrated along a narrow coastal strip and consists of Brunei Malays (66%), Chinese (11%), other indigenous groups (3%), with the remainder (20%) comprising other Borneo-indigenous groups such as Iban, and a still substantial number of expatriate workers who are temporary residents. Brunei's core national philosophy and ideology is *Melayu Islam Beraja* (Malay Islamic Monarchy). The Malay component refers to the official language, *Bahasa Melayu* (standard Malay), as designated in the 1959 Constitution, although the main lingua franca and the default language for everyday communication is the distinctive Brunei variety of Malay. Since 1985, 1 year after the resumption of full independence, Brunei has had a bilingual Malay and English language-in-education policy, with some subjects taught through the medium of (standard) Malay, and others, including Science and Mathematics, taught in English-medium. Under the current *Sistem Pendidikan Abad ke-21* ("Education System for the twenty-first century") English-medium operates right from pre-school through all levels (Haji-Othman et al., 2019). Hence most Bruneians under 35 years of age and educated to secondary-level or beyond are proficient in both standard and Brunei Malay, and in English (Goode, 2020).

The Context: Malaysia

Malaysia comprises the Malay Peninsula, bordering Thailand on the north and Singapore in the south, and two states on the island of Borneo, Sabah, and Sarawak. Based on the 2010 census, the total population of Malaysia was 28.3 million, with 20% of the population living in Sabah and Sarawak (dosm.gov.my). The ethnic breakdown in Malaysia is 67% Malays and indigenous groups, 25% Chinese, 7% Indians (7.3%), and 1% classified as others (e.g., Malaysians with Portuguese or Dutch ancestry). In Peninsular Malaysia, Malays make up 63% of the population (dosm.gov.my) while the indigenous groups, known as Orang Asli, comprise ∼0.7% of the population (https://www.jakoa.gov.my/data-terbuka-sektor-awam/). In East Malaysia, the Kadazan make up 26% of Sabah's population, while the Iban comprise 30.3% of the total population in Sarawak (dosm.gov.my). The language landscape in the public sector, including education, began to shift upon independence, which was first granted to the Malay Peninsula, Malaya, in 1957. Malay (*Bahasa Malaysia*) was declared as the national language (Article 152 of the Federal Constitution of Malaysia, 1963) as a means of promoting and creating a common national identity for the new nation which then comprised the Malay Peninsula, Sabah and Sarawak, and Singapore. Singapore left the Federation in 1965. At independence, there were schools with different mediums of instruction and curricula. Following independence, Malay began to replace English in government administration and public education. This process continued until the 1990s, when Science and Technology degree programmes, including Medicine, reverted to being taught in English-medium. In 2001 Malaysia reverted to English-medium for Science and Mathematics, then in 2009 this decision was reversed yet again, on mainly political, not educational grounds, and Malay-medium was reintroduced. Currently a "Dual Language Programme" is in place, allowing some schools a measure of choice as to medium of education for Science and Mathematics, including for public examinations. The changes in policy have done little to redress the rural-urban imbalance: in cities and large towns English functions as a second language; in rural areas English is effectively a foreign language, little used outside school classrooms.

Along with neighboring Southeast Asian nations Singapore and the Philippines, Brunei and Malaysia are categorized as "outer-circle" in Kachru's (2005) Three Circles model of world Englishes, since English has many intranational functions and both have distinct and well-described varieties of English, which are used between Bruneians and between Malaysians at varying levels of formality. It is beyond the limited scope of this article to give full descriptions of the linguistic and discoursal features of Brunei and Malaysian Englishes: for these, readers may refer to Deterding and Salbrina (2013) for Brunei English, and to Azirah and Tan (2012), among many other studies, for Malaysian English.

SOCIAL MEDIA IN BRUNEI AND MALAYSIA

Brunei is among the nations with the highest proportion of internet connection to its population with 95%. Of these 99% are active social media users (Kemp, 2021a). Among the most popular platforms are WhatsApp, Instagram and the Brunei Subreddit discussion forum. Wood (2016) highlights the importance of social media platforms for the development of Brunei English, owing to their popularity among younger bi- and multilingual Bruneians. This trend has become more pronounced since early 2020, when the COVID-19 pandemic first reached Brunei, necessitating lockdowns, restrictions on travel, and working and studying from home. This has led, in Brunei and Malaysia as elsewhere, to even greater reliance on social media. As "unregulated spaces" (Sebba, 2009), publicly available social media platforms offer an opportunity to examine emerging and shifting patterns of language choice and use.

Social media penetration and use in Malaysia is also high: there are 84.2% internet users and 86% social media users in the nation's population of 32.57 million (Kemp, 2021b). Whatsapp, Facebook and Instagram are among the most popular social media platforms in both countries.

THEORETICAL AND ANALYTICAL FRAMEWORKS ADOPTED

In the field of World Englishes (WEs), Kachru's (2005) interrelated notions of "nativization" of English and "englishization" of local languages are highly relevant to Brunei and to Malaysia. Kachru regards these as "two Janus-like faces of language contact situations involving English" (p. 135). These form the basis for discussion of Malay-English language contact in social media texts. Language contact affects the phonology, syntax, discourse and lexis of the languages concerned, both historically and even more so in the present time with the expanding affordances of social media. The insights of Makoni and Pennycook (2007, 2012) on disinventing and reconstituting languages, of Schneider (2014), on the evolutionary dynamics of WEs, and Schneider (2016) on hybrid Englishes, have influenced my views about WEs being by definition mixed codes (McLellan, 2020). Saraceni (2020, p. 716–717), notes that monolingual views of language and communication are challenged and lack relevance in Asian contexts, where boundaries between languages are fluid: fixed and bounded ideas of "native" languages, first, second and foreign languages are less applicable. According to Heryanto (2007, p. 43), the term "language" itself is not readily applicable in Southeast Asia, as it is not equivalent to the Malay term *bahasa*, which derives from the Sanskrit *bhāṣā* and has a broader connotational range comprising culture, politeness, upbringing and education.

With reference to the study of language use in social media, especially in multilingual contexts, Seargeant and Tagg (2014, p. 2) observe that "online social media are having a profound effect on the linguistic and communicative practices in which people engage, as well as the social groupings and networks they create."

Code Switching/Code Mixing/Language Alternation/Translanguaging?

Terminology in this field is a fraught and contentious area. It would be a fallacious oversimplification to claim that the terms are interchangeable, and arguing the merits and limitations of each would detract and distract from the main purpose of this article. Having used the term "language alternation" (henceforth LA) in McLellan (2005), I continue using it here, but am aware that translanguaging has gained credence in recent years and has outgrown its original domain, the analysis of teacher and student interaction in multilingual classroom contexts (Garcia and Li, 2014; Li, 2018, 2019).

With specific reference to multilingual online discussion forums such as Kytölä (2012) offers valuable insights on methods of both text selection and analysis, including the importance of going beyond the surface features of discussion forum texts such as LA and covering "naming (one's screen persona), heading (discussion topics), bracketing,….., slogans aphorisms, signatures" (Kytölä, 2012, p. 122).

As an analytical framework, the classification of social media texts into five categories, used by McLellan (2005), and

TABLE 1 | Language choice in two Brunei online discussion forums.

Posting text classification	McLellan (2005)		Deterding and Salbrina (2013)	
English-only (E–)	83	39%	41	40%
Main-language English (MLE)	36	17%	9	9%
Equal language alternation (=LA)	12	6%	5	5%
Main-language Malay (MLM)	57	27%	27	26%
Malay-only (M–)	23	11%	20	20%
Total	211		102	

subsequently by others (Deterding and Salbrina, 2013; 'Aqilah, 2020) is adopted for initial quantitative analysis. This aims to establish the frequency of LA as against monolingual English and monolingual Malay social media texts:

- monolingual English (E–)
- main language-English with some Malay (ML-E)
- equal language alternation of Malay and English (=LA)
- main language-Malay with some English (ML-M)
- monolingual Malay (M–).

It is axiomatic that monolingual texts are of equal interest and importance to texts which show a measure of LA. The temptation to label either monolingual texts or texts showing LA as "marked" or "unmarked" is therefore resisted, since the Bruneian and Malaysian text producers may choose any of these five, confident in the knowledge that they will be intelligible to their readers or interlocutors. **Table 1** shows the percentage of texts in the five categories in the corpus analyzed by McLellan (2005), and in a comparable corpus analyzed by Deterding and Salbrina, 2013, both corpora being collected from Brunei public discussion forums.

From the figures in Table 1, from McLellan (2005), there is a slight predominance of English over Malay as the choice for the main language, 56.4–37.9%. In terms of monolingual against mixed-language postings there is a near even split, 106 E- and M-, as against 105 showing some measure of alternation between languages. On the basis of the findings outlined in **Table 1**, the presence of some degree of LA is the norm for ML-Malay postings, whereas monolingual English is the norm for ML-English postings (McLellan, 2009). The figures obtained in the later study by Deterding and Salbrina (2013) demonstrate largely similar patterns.

Whilst lexical features of LA are perhaps the most evident surface feature for analysis using this framework, grammatical congruence and non-congruence between English and Sebba (1998) are also key aspects of any analysis of LA patterns. As discussed by McLellan (2009, p. 6), there are three major areas of morphosyntactic non-congruence:

- Noun phrase structure
- Pluralization of nouns
- Verb inflections.

English noun phrases have normally modifier-head constituent order ("my red car"), whilst in Malay, including Brunei Malay, the default order is head-modifier ("kereta merah saya"). In lines 1 and 3 of example (5) below, "abis bat" and "habis bat" ("dead battery") demonstrate the choice of English head-modifier order in the mixed noun phrase. In other instances the Malay order is found.

There is also non-congruence in how noun pluralization is marked: bound morpheme -s in English, reduplication of the noun in Malay (*bunga-bunga*, "flowers") if plurality is not evident from the context. Again, in mixed noun phrases either the English or the Malay pattern may occur: the English plural is retained on "others" in the mixed sentence in example (1)[1]:

(1) *I was the only person stay* di depan menaip *assignment*
                                   in front   type
     while the others   di ruang
                                   in level

     kedua
     second

     (free translation: "I was the only person staying in front typing my assignment while the others were on the second level," Brunei Subreddit, posted on 27th April, 2022).

But in (2), Malay reduplication is used for the English noun:

(2) ...seperti manuscript-manuscript atu
        like   DEM
     ("like those manuscripts," McLellan, 2009, p. 11)

Also found, though less often, is the use of the English -s plural suffix on Malay nouns.

Malay verbs, unlike English, do not have morphological suffix inflections marking tense, voice and aspect. Instead these are marked adverbially. Text (3) from is a mixed sentence in which "kana" is the passive voice marker.

(3) I was slowly kana angkat
                   PASS lift
     ("I was slowly lifted up," Brunei Subreddit, posted on 27th April, 2022).

This observes phrase structure rules of both English and Malay. Elsewhere in mixed verb phrases uninflected English verb base forms may be found, as in example (4):

---

[1]Example texts: layout and interlinear glossing conventions:
M3, F5, F6, F2: codes for male and female Whatsapp chat and group participants
Italics, English.
Normal font Malay.
ABBR, abbreviation.
Ar., Arabic.
DM, discourse marker.
1, first person pronoun.
3, third person pronoun.
DEM, demonstrative.
NEG, negation.
PASS, passive.
POSS, possessive.
RDP, reduplication.
REL, relative.
SG, singular.

(4) ...ex minister   atu,   kana *remove* from office due to this
                      DEM   PASS

     housing scheme"
     ("that ex minister was removed from office due to this housing scheme," McLellan, 2009, p. 13)

This shows dominance of Malay syntax through the non-appearance of the English -ed past participle marker.

These examples of morphosyntactic congruence and non-congruence show how non-congruence can be resolved by bilingual mixed text producers, illustrating the interface and interaction between Malay and English in social media texts.

ANALYSIS AND DISCUSSION OF EXAMPLES

Brunei: Whatsapp

'Aqilah (2020) compares WhatsApp two-party chats with multi-party groups, showing that patterns of alternation between Brunei Malay and English are similar to those found in face-to-face spoken interaction. She received the informed consent of all chat and group participants, and found a predominance of monolingual messages in both chats (64.8%) and in groups (73.2%), but a significant minority of messages with some LA. Example (5) is from a group chat between four participants:

(5) M3: Abis *bat*   kah?   *Or* rusak
         Finish ABBR-battery DM-question-tag broken
            ("Is the battery dead? Or is it broken?")
     F5: Mcm   rosak   *My* Adik   ckp
         ABBR-like broken   younger sibling ABBR-said
            ("It's sort of broken. My sister said")
     F6: [photo] *Dunno if* habis *batt or*   rosak
                            finish ABBR-battery broken
            ("Dunno if the battery is dead or broken")
     F2: *I think mine* tinggal di *uk* pasal   rosak   sudah
                        left in   UK   because   broken   already
            ("I think mine's left in the UK because it was broken already")

Whatsapp chats are asynchronous, but 'Aqilah is able to establish, through referring to time stamps, that this and similar interactions occurred with minimal intervals between turns, hence the group chat resembles face-to-face conversation. Intrasentential LA is evident in all the four messages, in spite of their brevity, and both languages contribute to the grammar and to the meaning. The conversation is designated = LA, with 13 words of Malay and 11 English. This short text serves as an example of Brunei English in a social media context. It shows informal features known to occur across social media worldwide, namely sentence fragments and abbreviations in both Malay ("mcm," macam, like; "ckp," cakap, say) and English ('Dunno'). The repeated Malay adjective occurs once with Brunei Malay spelling ('rusak'), reflecting Bruneian pronunciation, and three times with standard Malay spelling ('rosak'). Likewise there is one token of 'abis' (Brunei Malay) and one of the standard form 'habis' (finish(ed)).

Brunei: The Brunei Subreddit

The Brunei Subreddit (https://www.reddit.com/r/Brunei/) is a public discussion forum. It is available for all to read, but requires a username and password for those posting messages. As of 3 June 2022 it has 43,000 members, 8% of the nation's total population. Formerly the community (site managers) expected postings to be in English in the "random" discussion threads opened three times per week, but there has been an observable move toward higher frequency of use of Brunei Malay and of Malay-English LA in the past year, although the majority of postings are still in English-only. Every Monday, though, a thread "*Lakastah bekurapak dalam Bahasa Melayu pada hari Isnin*" ('Let's talk in Malay on Mondays') is opened. The use of Malay is strictly observed in this thread, in line with the warning (in English), "This is a thread to practice your Malay language, and posts not following this format will be removed/downvoted."

In code-mixed Brunei social media texts, the predominant pattern is standardized English alternating with Brunei Malay, as demonstrated in this example (6) from a posting in the Brunei Subreddit from 2019.

(6) *Old habits die hard* nya   urang, tapi *it gets worse* kalau
                        3POSS   person but   if
     dorang ani merasa inda di   terima lagi.
     3P   DEM feel   NEG   PASS   accept again
     ('Old habits die hard for these people, but it gets worse
     if they feel they are not accepted')

Text (6) is mainly Malay (ten words), with seven words of English. It is classified as equal language alternation (=LA), since both languages contribute to both the grammar and the lexical content.

Thus, although Bruneians predominantly choose to use English in social media, this is not to the exclusion of Brunei Malay. Hayani Nazurah (2021) finds that, in contrast to her corpus of Brunei Subreddit postings where English-only predominates, her corpus of 59 postings from the Brunei FM Facebook forum are 59% main-language Malay (ML-M) and 23% monolingual Malay (M-).

Table 2 shows an analysis of the thread from which example text (2) is taken, a discussion on the topic of employment of former prisoners.

The thread consists of eight postings, with four of the five language categories represented. Postings 4 and 7 are in monolingual English (E-) with no Malay, and posting 8 is in monolingual Malay (M–). The other five are in main-language Malay (M+) with some English insertions. The eight postings are from six different posters, as is evident from their Reddit nicknames (not given here for ethical reasons). One of those who posted twice used the same language choice, M+, for postings 1 and 5; the other used M+ in posting 3 and M– in posting 8. However, the English lexemes in posting 3, as noted in **Table 2**, are questionable: "program" is frequently used in Malay; "bebisnes" can be viewed as a case of intra-word mixing, a bicodal word, as it has the Malay actor-focus prefix "be-" with the English-derived root word "bisnes" (business), showing modification to comply with Malay orthography, which

is phonemic. The issue of words crossing from one language to another is discussed below in the section on englishization.

The analysis of threaded discussion forum postings, in which the participants use different language choices, demonstrates that they are not constrained in the choices they make, as they are aware that any of the five shown in **Tables 1**, **2** are fully intelligible. But the postings with LA will not be intelligible to anyone not proficient in Brunei Malay, hence although the Brunei Subreddit is open-access and free for anyone world-wide to read, few if any non-Bruneians are part of the online community who read this Subreddit, and even fewer non-Bruneians post messages on it.

Examples (7) and (8) are taken from Brunei Subreddit postings in Brunei English showing LA. (7) is classified as main-language English, E+:

(7) *Read more BM   story books and try to include*
     ABBR-language Malay
     peribahasa *in your* karangan. *The examiner*
     proverbs   essay
     would go "Wow this kid included peribahasa *in
     his* karangan?!!"
     ('Read more Malay language story books and try to include
     proverbs in your essay. The examiner would go "Wow this
     kid included proverbs in his essay"?!!')

(8) *so sometimes* ada   *vendors* cari   siapa mau
                    have   look for who   want
     share yr booth."
     ("so sometimes there are vendors who look for whoever
     wants to share your booth")

Example (7) here demonstrates LA in single Malay nouns within an English grammatical frame, whereas in (8), as in example (6), both Malay and English contribute to both the grammar and to the lexis, and this text is therefore =LA.

In example (9), from Hayani Nazurah (2021), the English phrasal verb 'give up' is written as "gibap":

(9) Kan *gibap* jualah   mula2,   tp
     DM   give up also-DM   RDP-start, ABBR-but
     apa   buleh buat.
     what can   do
     Those hardships pays later for sure.
     ("you can give up also at the start, but what can you do?....")

Pronunciation spellings such as "gibap" here, and also "bisnes," discussed above and in section analysis and discussion of examples below, occur often, and demonstrate the affinity of social media text with spoken interaction, in which such forms also occur.

Brunei: Instagram

According to Kemp (2021a), Instagram (IG) is the fifth most used social media platform globally, and the platform's potential advertising audience is 70.5% of the total population of Brunei aged 13 or over. Studies have been conducted of language use on IG by government ministries: Muhammad Nabil (2021) analyzed IG use by three Brunei government ministries, Education

Posting no.	Word-count	Language category	Malay words	English words	Notes
1	167	M+	149	18	Extract in example (2) above
2	59	M+	59	9	*uncle* ku….jangan kan *rely* arah *program* saja
3	14	M+	14	1	*program*…. "bebisnes" *(= Malay? Mixed = Bicodal?)*
4	78	E-	0	78	Std. E: *To be fair when a person has been jailed…*
5	70	M+	65	5	*Yeap*, banar tu cakap kita. *and etc*. x2
6	18	M+	17	1	Ada jua *program* utk dorg
7	27	E–	0	27	Std. E: *maybe the family can start up a business for him*
8	12	M–	12	0	…sama ada ya banar banar atau inda

(Islamic), Religious Affairs and Health, in terms of their language use, including responses by the Brunei public. For Religious Affairs all the IG communications are announcements in Malay or Arabic, and there are no responses. The Education and Health Ministries use both Malay and English for announcements and media releases, and both languages are found in responses, reactions and queries to these. Since the onset of the Covid-19 pandemic the Ministry of Health's IG communications have taken on greatly increased importance, with daily briefings and media conferences chaired by the Minister of Health along with other ministers, at which both Malay and English are used, often with LA.

In a study of commercial advertising and customer responses on IG in Brunei, Diyanah Maimunah (2021) found that English was the preferred language by advertisers, although some included LA or were in Malay only. In an online survey she conducted, young Bruneian IG users aged 18–26 reported a preference for mixed Malay and English when responding to IG advertisements, assuming that any of the five code choices would be acceptable and effective. She gives an example of an E+ advertisement with only one word of Malay by Stack, a local burger restaurant, which generated a higher number of responses in Brunei Malay than in English. This is an exception to the more frequent pattern of customers responding use the same code choice among the five as in the initial advertisement.

Brunei: Facebook Status Updates

Nurdiyana and McLellan (2016) analyzed a corpus of 239 Facebook status updates by Bruneians, with their informed consent. 8.8% of these were in Malay only, 60.3% were in English only, and 25.5% had some mixing of English and Malay. The figures do not add up to 100% because Arabic, frequently mixed with Malay by Bruneians, was also part of the analysis. Some examples from this corpus follow:

(10)   (intrasentential, switch at phrase boundary)
   Black out n   aku   kepanasannnn!!!!!!!!!
                    1SG   hotnessssss
   ('Black out n I'm hot!')

(11)   (intersentential, switch at phrase boundary)
   mentally & physically tired…   malas   ku   malas!!
                                      lazy   1SG   lazy
   ("I am mentally and physically tired! Lazy I am lazy")

Both these major syntactic LA strategies signal high levels of bilingual Malay-English proficiency, including the ability to alternate without breaking the syntactic constraints of either language (Muysken, 2000, p. 122). Example (12) shows rich intrasentential mixing, whilst (13) is a trilingual posting:

(12)   just finished one class dan dua lagi   coming ohhh
                                and two more

   letihnya
   tired-POSS
   ("I've just finished one class and two more to come, ohhh I'm tired")

(13)   Assalamualaikum   pasal   replacement intro to
   Peace be upon you (Ar.) about

   teaching, if on monday okay kh?
                                DM

   If okay, what tym bisai?
                  time good

   ('Peace be upon you…about the replacement Intro to Teaching (class), if it's on Monday is that okay? If okay, what time is good?)

This trilingual example (13) begins with the Arabic greeting, followed by the Malay conjunction "pasal" governing an English noun phrase. The first sentence ends with the Malay question tag "kh" (short for "kah?"). The second sentence reverts to English prior to a further switch to Malay for the adjective "bisai." In line with the grammar of Malay there is no copular verb in either sentence: as shown in the free translation, this would be required in standard English.

These examples from Facebook status updates provide further evidence of the diverse patterns of LA in this social media, again demonstrating that all the five code choices, monolingual or mixed, are acceptable and used.

Malaysia: Online News Media Reader Responses

This section is adapted from McLellan (2016). The print media in West Malaysia publishes monolingual Bahasa Malaysia, English, Chinese and Tamil newspapers, but the Malaysian Borneo states of Sabah and Sarawak have newspapers which are bilingual such as the *Utusan Borneo*, published in Sarawak in Bahasa Malaysia and in the indigenous Iban language (once weekly). In Sabah the trilingual *New Sabah Times* has sections in English, Bahasa

Malaysia, and in the indigenous Kadazandusun language. All are available online as well as in print versions.

One major affordance of online editions of newspapers is reader response, where readers are invited to post their responses to the reports, editorials and feature texts. These have the same interactional features as threaded online discussion forums, and one report can provoke a series of responses akin to a face-to-face conversation. They are moderated by the newspaper staff, and responses deemed inappropriate, offensive or potentially libelous are deleted.

A small corpus of 18 texts with 58 reader responses was collected from the *Utusan Borneo* and the *Borneo Post* newspapers between September 2013 and June 2014. These were analyzed in terms of markedness, defined in this context as instances where the language choice of the reader response differs from that of the original posting. An unmarked response is one which uses the same language as the newspaper report. **Table 3** gives basic information about this corpus.

Almost one-third of the reader responses show a language choice which differs from the news report. Text (14) is an example where the response text is code-mixed Malay and English, whilst the original *Borneo Post* report was in English.

(14) Headline of English news report: Include Dayaks in MEB – Masing
Date: 16 September 2013
Reader response text 1 (of 1)
YB...macam mana mau buat "*road show*" kalau "*road*" nya tiada. Buat "*boat show*" lah tauke...sambil2 tu buat "*miring*" sekali...untuk menyejukan hati antu yang selalu halang org buat jalan ke kapit
(*Free translation*: Yang Behormat ("The Honorable" - honorific for elected representatives), how can you do a "road show" if there is no road? Why not do a "boat show," boss?... and then a "miring" ("Iban traditional ceremony") as well, to freeze the souls of the ghosts which have always prevented people from making a road to Kapit.)

This reader response is in main-language Malay, with three English noun-phrase insertions, all flagged with inverted commas, and one in Iban, "miring" (Iban blessing ceremony).

Example (15) shows a response in English to a report of a road accident in the Iban-language section of *Utusan Borneo*.

(15) Headline of news report: Lelaki parai dalam kes bebadi jalai alun
                    (Iban: "Man injured in road accident case")
Date: 21 September 2013
Reader response text
There are hundreds of fatal accident ramped against planted trees along Kidurong road since some decades ago. It would be safer for the road without big trees all along it even if the road is hot without shade. Yes there is a beauty of the scene but people's life is more important than that. Bintulu town planner should look seriously into this matter. I recommend flowers and small species type plant may be planted to beautify the road side. Do not blame the road users alone for careless driving. No one is expecting an accident.

TABLE 3 | Language choice in 58 reader response texts in the Borneo Post and *Utusan Borneo*, Sarawak, Malaysia.

	Number of responses	%
Unmarked	39	67.2
Marked	19	32.8

This example shows a higher level of syntactic complexity and lexical density than the response text (14) above, hence it is more formal. Reader response texts can have variable levels of formality, because the asynchronous format of this online genre permits readers to take time in planning their texts before posting them.

As with the Brunei Subreddit discussion forum, Instagram and Facebook texts, some of the reader response texts demonstrate LA between Malay and English, with Iban as a further possible language choice. Example (16) below is one of nineteen responses to a *Borneo Post* report. Of these, three show Malay/English LA and another is in monolingual Malay.

(16) Headline of English-language news report: See says difficult for Taib to retire
Date: 10 September 2013
Reader response text 4 (of 19) – extract
And one thing that irks me most was when the teachers from the Peninsular Malaysia promotes us about the KLIA airport and all those modern infrastructure stuffs. Perghh! Apa diaorang ingat kita Sarawak*ian* masih tinggal kat hutan kah? *Please* lah beb, Sarawak dah jauh maju dah dari pemikiran diaorg.
And actually, we Sarawakians has a lot of young people who has tons of great potential on becoming a great teacher. But *shrugs* apa boleh buat, orang mahu hantar banyak cikgu Semenanjung ke Sarawak.
(*Free translation of Malay in this text*: Do they think we Sarawakians still live in the jungle? Please, friend, Sarawak is far more developed than people think. …what can we do, they want to send many teachers from Peninsular Malaysia to Sarawak.)

The pattern of LA in example (16) is intersentential E>M>E>M.

The style is less formal than the monolingual English response text in example (15), and is closer to the conversational discourse found in example (14). The language alternation is a contributing factor to the informality.

In instances where there is LA, or where the reader responses are in a different language to that of the original online news report, there is an assumption that readers share the same multilingual competences as the text producers, and that they will have no problem understanding the response texts. The texts which show a measure of LA thus demonstrate how English coexists and functions as a resource, alongside the other available languages, for the online text producers. These examples from Malaysian online media texts show similar patterns to those from the Brunei social media platforms: all the five possible choices in **Table 1** are available, and those responding to postings feel under no constraint to retain the same language choice as the original

TABLE 4 | Pronunciation spellings and other translingual forms.

Ice box = Isbuk	Confuse = Konpius
Action (lying) = Eksen	Low batt = Lubit
Sound = Saun	Facebook = Pisbuk
Round (walking in circle) = Raun	Local = Lokal
Motor car = Matuka	Support = Sapot
Mobile = Mubail	Parking = Paking
Charlie Chaplin (funny) = Cali	Go Start/Go Astern
Worry = Uri	(reverse) = Gustan
Current (electricity) = Karan	Television = Tilibisin
Time = Taim	Queueing = Bekiu
Drive = Driba	Fashion = Pisin
Driving = Dribin	Guarantee = Gerenti
Cute = Kiut	Challenge = Celen
Engine = Injin	Driveshaft = drepsap
Boring = Buring	Balance = Bilin
Style = Stail	F Off = Pakop
Confirm = Kompom	Colgate = kulgit
Company = Kompani	

Source: The Brunei subreddit, posted on 9 October 2018.

media report or forum posting, as they know that their responses will be intelligible to readers.

THE OTHER SIDE OF THE COIN: ENGLISHIZATION OF MALAY

As noted in the introduction, language contact works bidirectionally, and there is ample evidence in Malay texts on social media of the influence of English, leading to englishized or anglicized Malay as an area of investigation. In addition to examples discussed above, Hayani Nazurah (2021) notes the of "tym" or "tyme," deriving from English "time," but used as a conjunction to mean "when" in past-tense narratives in a Brunei Subreddit text. Hence this is a shift of word-class. She also notes examples of pronunciation spellings "sinsas" (census) and "mudipait" (modified). A Brunei Subreddit posting from 2018 gives a longer list of mostly pronunciation spellings and other translingual forms used creatively by Bruneians when speaking or interacting in social media in Brunei Malay in **Table 4**.

One morphosyntactic feature of englishization involves Malay affixes attached to English root words in ML-M texts, as is the case for "berbisnes" mentioned in section analysis and discussion of examples. Example (17), from the Brunei Subreddit (25 April 2022) shows the Malay prefix "be"- affixed to the originally English "disiplin":

(17) …untuk menolong orang yang inda *bedisiplin*
       dari to   help   person REL NEG disciplined
       from segi kewangan
       sector   financial
       ("to help people who are undisciplined from a
       financial perspective")

Example (18), also from the Brunei Subreddit (25 April 2022), shows the Malay possessive suffix "-ku" affixed to the originally English noun "boss."

(18) atu   na   pandai abis   utang tu
     DEM NEG can   finish debt DEM
     bossku
     boss-1POSS
          ("they cannot clear their debts owed to the boss").

These demonstrate integration of originally English words into Brunei Malay, and can be seen as reshaping and repossession of the English words, as well as a blurring of boundaries between the two languages, a consequence of extended language contact.

CONCLUSION

There are two core questions arising from this investigation which can be addressed here, although perhaps not fully answered:

- Is it valid to posit the language of social media as a new variety comprising both local and global influences and inputs?

It is evident that the increased availability and popularity of social media is bringing about changes in patterns of language use. "Lol" (laugh out loud), for example, is now used as a finite verb taking -s, -ing and -ed suffixes. But from the examples presented and discussed here it is also evident that social media platforms are not pushing users toward the use of monolingual English. Patterns of mixing are found, many of which reflect those found in everyday informal conversation. The subsections under 4 and 5 above, when compared and contrasted, show varying patterns of lexical and morphosyntactic LA, and varying patterns of intra- and intersentential mixing. So it may be more reasonable to posit a range of new varieties, not just a single social media variety, which are developing within the diverse social media platforms in Brunei and Malaysia.

- Is social media driving changes in varieties of World Englishes in Southeast Asia?

The evidence assembled here suggests that this question could be answered in the affirmative, provided that there is acceptance of the premiss of World Englishes being by definition code-mixed varieties demonstrating features of the other languages in the multilingual ecologies of nations such as Brunei and Malaysia where distinct varieties of English have developed. Even in monolingual English texts, variable features such as countability of nouns ("an advice," "equipments," "furnitures") may occur, caused in part by the influence of Malay (McLellan, 2020, p. 428–430). However, in spite of the evidence of the wide spread and increasing importance of social media in both Brunei and Malaysia, discussed above, the answer to this major question must remain tentative, owing to the synchronic nature of the data collected and discussed in this paper. Further research into social media language choice and contact patterns could be conducted using longitudinal data collection methods and more extensive textual corpora.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

'Aqilah, A. (2020). *Language Choice and Code Switching in Brunei in Synchronous Electronically Mediated Communication*. PhD thesis, FASS, Universiti Brunei Darussalam.

Azirah, H., and Tan, R. (2012). "Malaysian English," in *Englishes in South East Asia: Features, Policy and Language in Use*, eds E. L. Low and A. Hashim (Amsterdam: John Benjamins), 55–74. doi: 10.1075/veaw.g42.07has

Deterding, D., and Salbrina H. S. (2013). *Brunei English: A New Variety in a Multilingual Society*. Dordrecht, Netherlands: Springer Verlag.

Diyanah Maimunah, K. (2021). *Language Patterns of Online Advertisements and Responses on Bruneian Instagram (Final-Year Research Report)*. Universiti Brunei Darussalam, Brunei Darussalam.

Garcia, O., and Li, W. (2014). *Translanguaging: Language, Bilingualism and Education*. Basingstoke: Palgrave Macmillan.

Goode, C. (2020). English language in Brunei: use, policy, and status in education: a review. *Indonesian J. English Language Teach.* 15, 21–46. doi: 10.25170/ijelt.v15i1.1411

Haji-Othman, N. A. (2012). "Is it always English? 'Duelling aunties' in Brunei Darussalam," in *English Language as Hydra: Its Impact on Non-English Language Cultures*, eds V. Rapatahana and P. Bunce (Clevedon: Multilingual Matters), 175–190. doi: 10.21832/9781847697516-016

Haji-Othman, N. A., and McLellan, J. (2014). English in Brunei. *World Englishes* 33, 486–497. doi: 10.1111/weng.12109

Haji-Othman, N. A., McLellan, J., and Jones, G. M. (2019). "Language policy and practice in Brunei Darussalam," in *The Routledge International Handbook of Language Education Policy in Asia*, eds A. Kirkpatrick and A. Liddicoat (London: Routledge), 314–325. doi: 10.4324/9781315666235-22

Hayani Nazurah, Y. H. (2021). *Code-Switching in Online Interactions: Comparative Study of Facebook and Reddit Comments by Bruneian Users (Final-Year Research Report)*. Universiti Brunei Darussalam, Brunei Darussalam.

Heryanto, A. (2007). "Then there were languages: Bahasa Indonesia was one among many," in *Disinventing and Reconstituting Languages,* eds S. Makoni and A. Pennycook (Clevedon: Multilingual Matters), 42–61.

Kachru, B. (2005). *Asian Englishes: Beyond the Canon*. Hong Kong: University of Hong Kong Press.

Kemp, S. (2021a). *Digital 2021: Brunei Darussalam – Data Reportal*. Available online at: https://datareportal.com/reports/digital-2021-brunei-darussalam (accessed November 3, 2021).

Kemp, S. (2021b). *Digital 2021: Malaysia – Data Reportal*. Available online at: https://datareportal.com/reports/digital-2021-malaysia (accessed November 5, 2021).

Kytölä, S. (2012). "Multilingual web discussion forums: theoretical, practical and methodological issues," in *Language Mixing and Code-Switching in Writing: Approaches to Mixed-Language Written Discourse*, eds M. Sebba, S. Mahootian and C. Jonsson (Abingdon: Routledge), 106–127.

Li, W. (2018). *Translanguaging and Code-Switching: What's the Difference?* OUP Blog. Avaialble online at: https://blog.oup.com/2018/05/translanguaging-code-switching-difference/?fbclid=IwAR2DZVo5F7H6xSpzZIPLD4i2yKHXFcBjK_xZsqbDSItD2AuxqMlWNSynFDY (accessed November 5, 20210).

Li, W. (2019). Translanguaging as a practical theory of language. *Applied Linguistics* 39, 9–30. doi: 10.1093/applin/amx039

Makoni, S., and Pennycook, A. (2007). *Disinventing and Reconstituting Languages*. Clevedon: Multilingual Matters.

McLellan, J. (2005). *Malay-English Language Alternation in Two Brunei Darussalam Online Discussion Forums*. Ph.D. thesis, Curtin University of Technology, Australia.

McLellan, J. (2009). When two grammars coincide: Malay-English code-switching in public on-line discussion forum texts. *Int. Sociol. Association RC25 Newsletter* 5, 1–22.

McLellan, J. (2016). "English and other languages in the online discourse of East Malaysians," in *English in Malaysia: Current Use and Status*, eds T. Yamaguchi and D. Deterding (Leiden: Brill), 125–146.

McLellan, J. (2020). "2ⁿᵈ edn Mixed codes and varieties of English," in *The Routledge Handbook of World Englishes*, eds T. A. Kirkpatrick (Abingdon: Routledge), 425–441. doi: 10.4324/9781003128755-28

Muhammad Nabil, K. H. (2021). *Examining Language Use by the Ministry of Education, Ministry of Health and Ministry of Religious Affairs on Instagram (Final-Year Research Report)*. Universiti Brunei Darussalam, Brunei Darussalam.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Nurdiyana, D., and McLellan, J. (2016). Gender and code choice in Bruneian Facebook status updates. *World Englishes* 35, 571–586. doi: 10.1111/weng.12227

Saraceni, M. (2020). "Globalization and Asian Englishes," in *The Handbook of Asian Englishes*, eds K. Bolton, W. Botha and A. Kirkpatrick (Hoboken, NJ: Wiley Blackwell), 707–724. doi: 10.1002/9781118791882.ch31

Schneider, E. W. (2014). New reflections on the evolutionary dynamics of World Englishes. *World Englishes* 33, 9–32. doi: 10.1111/weng.12069

Schneider, E. W. (2016). Hybrid Englishes: an exploratory survey. *World Englishes* 35, 1–16. doi: 10.1111/weng.12204

Seargeant, P., and Tagg, C. (eds.) (2014). *The Language of Social Media: Identity and Community on the Internet*. NewYork: Palgrave Macmillan.

Sebba, M. (1998). A congruence approach to the syntax of code switching. *Int. J. Bilingualism* 2, 1–19. doi: 10.1177/136700699800200101

Sebba, M. (2009). *Unregulated Spaces: Exploring the Potential of the Linguistic Margins*. Available online at: http://www.ling.lancs.ac.uk/staff/mark/unreg.pdf (accessed October 21, 2015).

Wood, A. S. (2016). "The iscourse of online texts in Brunei: extending Bruneian English," in *The Use and Status of Language in Brunei*, eds N. Azam Haji-Othman, J. McLellan and D. Deterding (Dordrecht: Springer), 187–200.

Contact-induced grammar formation: A model from a study on Hiberno-English

Tamami Shimada*

Faculty of Languages and Cultures, Meikai University, Urayasu, Japan

This article examines how certain characteristic grammatical forms in Hiberno-English (HE) are the result of a dynamic process of language formation guided by language contact. A first language contact between Irish and English gives rise to the grammar formation of HE, and a second contact between HE and other varieties of English, presumably over the past 50 years or so, has pruned HE to fit the speakers' awareness toward the standard norm. Examinations of the expressions of tense/aspect and information structure in HE lead to suggestion of grammatical oppositions being inherited from Irish and the resilience of this inheritance in present-day HE. Taking three salient characteristics of HE, the *be after* perfect, the *do be* habitual, and the *'tis.....* construction as windows to its underlying properties, the article surveys earlier forms in the rise of HE and describes some facets of contemporary HE. One of the central issues in the examination is Irish language traits and their realization in HE morphosyntax. The article concludes by proposing an integrated perspective across the characteristics and a model to capture the grammar formation of HE, which can be applied to find similarities and contrasts with other language contact phenomena.

KEYWORDS

language contact, Irish English, *do be* habitual, *after* perfect, *'tis*, *it*-cleft, information structure, language change

Introduction

In Ireland, like many other places, various forms of English have penetrated the everyday lives of the population *via* the media. Within this state of affairs, Hiberno-English (HE), the English spoken by the Irish, is on a path toward convergence with other varieties, but it still preserves some of its Irish heritage. This article addresses this heritage in its expressions for time and its manner of expressing information structure, two elements which form part of the core of the grammar. It discusses how HE has been formed and developed up until today by an examination of three salient grammatical features. The article focuses on varieties of Southwest Hiberno-English (SwHE), spoken in the counties of Cork and Kerry where a language shift from Irish to English has not progressed to the same extent as in other regions of Ireland. In the investigation of contact-induced properties of grammar, SwHE is a prime subject due to its robust connection with the Irish language.

The article investigates the process behind the establishment of constructions in present-day HE. One of our concerns is the process with which the *be after* perfect, as exemplified in (1), has gained its present form and usage.

(1) We are after missing the bus.
      "We have just missed the bus". [Cork City]

(1) is an example of *be after V-ing* from contemporary varieties of SwHE. The speaker said this when "we" were running to catch a bus leaving from the bus stop in front of us. The basic property of *be afterV-ing* is to denote that, as of the speech time, a certain activity or event has been completed. The *be after* perfect is often labeled as the "hot-news perfect". Using *be after*, the speaker presents the event as "hot-news" and signals their emotional attachment to it.

Besides the *be after* perfect, this article highlights key distinctive constructions (*vis-à-vis* Standard forms of English) including the *do be* habitual and the *'tis.....* pattern. (2) is an example of the *do be* form in SwHE. The *do be* form, generally speaking, seems to be falling out of use if we look at the contemporary situation and speakers' attitudes toward the form. However, the majority of HE speakers, both urban and rural, even including the younger-middle generations, have certain knowledge of this construction (Shimada, 2016a).

(2) He does be eating in John B's.
      "He usually eats in John B's (pub)". [Co. Kerry, elicitation]

In SwHE, *do be (V-ing, AdjP)* is used to describe an inherent property or a habitual behavior of the agent, which is expressed in the *-ing* form after *do be*. Thus, it is called habitual *do be*, with the *do* being unemphatic. Formal realization of the habitual categories varies in dialects of HE; northern varieties tend to have *be(z)∼do(se) + V,* while speakers of southern varieties use *do be* (Henry, 1957; Harris, 1986).

The *'tis.....* construction is the third salient characteristic of HE that we highlight in our investigation of contact-induced grammar formation. Although this construction displays similarities in appearance to the *it*-cleft in standard varieties of English (StE), clearly, important differences have been pointed out (Filppula, 1999; Shimada, 2018).

The *'tis....* pattern in SwHE is characterized by its high frequency of use, the wider variety of phrasal categories allowed in the supposed "focus" position, and the primacy of *that*-lessness (see The *'tis...* construction). (3) is an example of the *'tis.....* construction.

(3) I suppose 'tis boozing on brandy you are with McFillen.
      "I suppose you are boozing on brandy with McFillen".
      (prosodic prominence on *boozing*) [John B. Keane STD56, = (29)]

(3) admits a different context of the use of *'tis...* in HE from that of the *it*-cleft in StE. In (3), *'tis* presents a proposition that

is characterized as a supposition by *suppose*. In the clause of *'tis,* a marked constituent order results; the fronted constituent *boozing on brandy* is salient in the clause *you are boozing on brandy with McFillen*. The sentence expresses that this particular component of the state of affairs is highlighted in the speaker's mind. The *'tis...* pattern in HE, despite an apparent resemblance to the *it*-cleft in general English, exhibits a different information structure decoded in the sentence.

(1)-(3) are present-day forms in SwHE. The *be after* perfect and the *do be* habitual are expressions of time, while the *'tis.....* construction is concerned with information structure. Using these salient characteristics as the basis of discussion[1], we will investigate the central part of contact-induced grammatical formation.

Studies on features of HE that are distinctive from StE including the seminal studies of Bliss (1979), Harris (1986, 1993), Filppula (1999), Siemund (2004), Hickey (2007), and Kallen (1997, 2013) and (Kallen and Kirk, 2007) have made valuable contributions to the description of HE and the superstratal-vs.-substratal debate. Recently, development of corpora such as ICE-Ireland (Kallen and Kirk, 2008) and CORIECOR (McCafferty and Amador-Moreno, 2012), which contains letters of immigrants, improve our understanding of HE, both its current and earlier forms. Today, research on Irish English particularly flourishes in areas of discourse markers, as represented by the volumes edited by Barron and Schneider (2005), Amador-Moreno et al. (2015), and in sociolinguistic interests (Hickey, 2016; Hickey and Amador-Moreno, 2020). Another important research stand has been on the connection of the grammar of Irish English with other Celtic Englishes, which has been elaborated by the authors of *The Celtic Englishes,* edited by Tristram (1997). Also important to note is the pursuit for how Celtic languages have influenced English (Filppula et al., 2008).

Many past studies concerning the grammar of Irish English have taken a form-to-form comparative approach, focusing on the way in which forms in HE are deviant from StE and which forms in Irish or earlier English they most closely correspond to. This study, taking a different angle, examines the systematicity of the connection between HE and Irish. It seeks to find general principles from which features of HE result naturally. Taking the three distinctive features above as windows into the underlying mechanisms of the formation of HE, this study examines how their current forms and meanings are the result of a dynamic process of language formation guided by language contact. It will survey earlier forms in the rise of HE (section Earlier examples

---

1 The phrase "salient characteristics" is used when comparing with other varieties of English including British Standard English. The idea of the "salient features"/"salient characteristics" comes from Filppula (1999), who uses the phrase to refer to features or characteristics of HE *vis-à-vis* Standard English. Such salient characteristics help to uncover influence from the Irish language and, as in the current study, can serve as "windows" to investigate contact-induced properties of grammar.

and the form-function establishment) and describe some details of contemporary HE (section Irishness realized in English). This article will finally introduce an integrated perspective across the characteristics to discuss how elements from Irish have been inherited and realized in English, along with how these realizations have been updated until today[2].

Data and methodology

For the investigation of present-day SwHE, this study employs a fieldwork-based qualitative methodology, which includes elicitation techniques, participant observation, and interviews, both linguistic and sociolinguistic. The majority of the author's data come from regular stays in Cork since 2002 and Listowel since 2003. The author has also visited Dublin and an Irish-speaking district called Dingle Peninsula. The examples of forms used in present-day SwHE include both the ones the author encountered in natural speech and email texts and the ones that were elicited during linguistic and sociolinguistic interviews. Furthermore, as part of the analysis of the process of the formation of HE, the author draws on examples and studies concerning earlier varieties in the literature review.

The elicited data used in this article come from five consultants living in Cork City and County Kerry whose birthdates range from the 1920s to the 1970s. The linguistic interviews were semi-structured with questions inspired by hypotheses the author drew based on a corpus of John B Keane's playscripts and letter series written mainly in the 1960s and 1970s (refer also to note 8 for the "Keane corpus"). The author had compiled this corpus for previous research and had selected sentences in his studies, which included target features including the three addressed in this article: *do be*, *be after*, and the *'tis…* construction. Analyzing this corpus with guidance from the pioneering descriptions of HE made by Filppula (1999), among others, led to construction of hypotheses concerning morphosyntactic environments and contexts in which the features appear. The HE variety reflected there is most likely a "stative HE" used during the 1960s and 1970s, which had not yet been exposed to the dominant varieties of English made commonly available in Ireland by TV and the Internet (refer to Two different contacts in the formation of present-day HE and Figure 1 for the transition of HE in timeline). This study initially started with the author's observations of contemporary HE in use in terms of Filppula's (1999) "distinctive features",

_____

2 The article is based on the author's prior studies and descriptions, which include Shimada (2018) for the analysis of the *'tis…..* construction, Shimada (2013) for *do be* habitual, and Shimada (2008, text in Japanese) for the *be-after* perfect. The aim of this study is to integrate properties of these features and their formation into a theory of contact-induced grammar formation.

examining what led to the robustness of some features and the decline of others.

Formation of southwest Hiberno-English in a sociohistorical timeline

A language is naturally affected by the situation of its community. Thus, HE has changed along with the changing times of its community. In the following, after first covering the flow of history with a focus on language, we will consider the formation of HE and its changes over time.

Formation and the development of SwHE

In the southwest of Ireland, including the counties Cork and Kerry, statistics of Irish and bilingual speakers in the 18th and 19th centuries show that Irish had not succumbed to English before the Famine in 1840s, but that it was not inherited by the generation of speakers who were born post-Famine. According to FitzGerald (1984), the 1881 census indicates that 41% of the population aged 70–79 in 1881 had been Irish-speaking in youth, and 77% in Munster, the southern province (ibid: 125). From his examination, it is concluded that Irish remained vigorously spoken in South Cork and North Kerry in 1841 and was more prominent in southeast and southwest Cork and Kerry from Tralee southwards (ibid: 128). Nevertheless, a large proportion of the Irish-speaking population may have already been bilingual; data illustrating that "Irish and English" speakers were outnumbered four times by "Irish-only" speakers in 1851 corroborate this (ibid: 140).

After the Great Famine, there was little room for doubt that the language shift progressed rapidly. In southwest Ireland, the shift started at a comparatively late date, understood to be around the mid to late 19th century. However, the earlier English-lexifier varieties emerged in most cases as the result of natural or untutored L2 acquisition. At the beginning of the 19th century, the only schooling available for the population was obtained from "hedge" schools (Edwards, 1981, p. 241), where local masters were paid a small sum by parents, and Irish was likely to be used as the language of instruction (Ó Cuív, 1986, p. 380, 381). High illiteracy figures show the scarcity of schooling in the mid to late 19th century. The illiteracy figure for the whole country was 53% in 1841 and 47% in 1851, but it decreased to 18% in 1891. For the counties Kerry and Cork, the focus of the author's investigation, it was 70 and 66%, respectively, in 1841 (Edwards, 1981, p. 242, 243). Odlin (1997, p. 12), using demographic data to assess the effect of seasonal migration, states that "most workers from Ulster went to Scotland, whereas the southwestern counties of Kerry, Cork, and Limerick sent few migrants at all to Britain; most workers from these areas sought

work in more prosperous Irish counties such as Wexford". Also, it is assumed that before the 19th century, the acquisition of English by many Irish speakers in southwest counties did not result directly from schooling.

Supposing that the formation of HE accelerated in some south-western communities in the latter half of the 19th century, we can estimate that sometime between 1930 and 1950 a largely stable system for HE had taken root in southwest Ireland. However, that stable system, amid the changes in environment brought by TV broadcasting, widespread education, and entry into the EU (European Union), along with emigration and migration, came into contact with other major forms of English, and it underwent standardization and leveling. Following this, the social situation brought by the rapid economic growth of the latter half of 1990s, called the Celtic Tiger, brought even further changes to the English spoken in Ireland. As many varieties of English, including StE, flowed into Ireland, the once stable system began to falter. HE is now experiencing what we might call a "second contact", this time with other forms of English.

Two different contacts in the formation of present-day HE

The linguistic history leading up to the formation of HE and subsequent changes to it can be roughly summarized as shown in Figure 1. Figure 1 is the author's summary of the historical background of the development of HE.

On the timeline of the development or transition of HE, two contacts in different times seem to be recognized as prominent milestones. The first contact is that of Irish and English. The second is that of HE and major varieties of English or what can be described as mainstream English. Thus, in a simple way, three stages can be distinguished: (I) English enters the Irish-speaking community, (II) HE forms and gains relative stability, and (III) HE undergoes contact and convergence with mainstream English. The dynamics of contemporary HE can be regarded as the outcome of two different occurrences of contacts, one completed and the other still in progress.

During the period in which English, in some form, came into common usage in Ireland, stage (I) in Figure 1, Irish monoglots became bilingual mostly through contact and interaction with other Irish people who had already acquired English. This English is most likely the prototype of HE. According to Bliss (1972, p. 63), in the mid-17th century, English was "acquired, gradually and with difficulty, by speakers of Irish; and in the process of their acquisition of it they modified it, both in pronunciation and in syntax, toward conformity with their own linguistic habits". There were, of course, different processes working at the same time in different regions and spurred by different contact situations, all of which fostered language shift and the formation of HE. If we see this phenomenon from a micro-perspective, that is, focusing on a speech community in one particular area and not on the whole country of Ireland, HE has been shaped in a somewhat similar way to other contact varieties which came into being as a result of the two different languages. Although the spread of the English language in the whole of Ireland has taken a long time, language shift at the level of each community is, in general, accomplished in three generations over approximately 100 years.

Earlier examples and the form-function establishment

This section addresses examples of the three constructions found in the texts written in the 16th-19th centuries to see initial forms of HE. In terms of the timeline given in Figure 1, stages (I) and (II) are discussed. The focus is on the *do be* and *be after* V-*ing* forms, concerning the expression of tense/aspect, and the '*tis...* construction, concerning the expression of information structure. For the *do be* and *be after* constructions, there is already a wealth of research, so the main points will be summarized as part of the literature review. For the '*tis...* construction, there seems to be a room for further research. Examples of the '*tis...* construction, mainly from Bliss (1979), will be analyzed to examine how the Irish language was mixed into English speech/writing in the early stage of HE.

Be after perfect

Examples of *be after* V-*ing* emerged from an earlier stage of the development in HE. The *be after* form is traced back to the 1680s, as reported by Filppula (1999, p. 103) and Kallen (1994, p. 173). The following is an example of *be after* V-*ing* cited from Bliss (1979, p. 133).

(4) Deare Catolicks, you shee here de cause dat *is after bringing* you to dis plaace: 'tis come bourying you are de crop, de cadaver, of a verie good woman, God knows!... [*Report of a Sermon* (1698)].

A common view in the literature on HE is that the *be after* construction is calqued on the Irish *tréis (∼tar éis)* "after" construction. This view is based on the obvious parallelism between Irish and HE, shown in (5) below, and the fact that no recoded form of English offers an alternative model (Greene, 1979, p. 126). The *be after* NP form is given in (5a) and *be after* V-*ing* in (5b).

(5) a. Tom is after his supper. [HE]
   Tá Tomás tréis a shuipéar. [Irish]
   be.pres. *Tomás after his supper*
   "Tom is in the state where he has had supper".

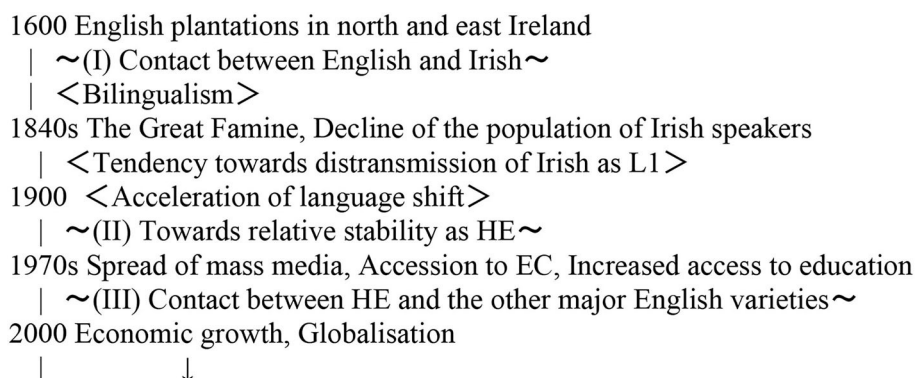   b. I am after taking three plates from the cupboard. [HE]

FIGURE 1
Timeline of the sociohistorical background of HE.

Tá mé tréis trí phláta a thógáint ón gcófra. [Irish]
be.pres. *I after three plates* PRT *take*.$_{VN}$ *from-the cupboard.*
"I've just taken three plates from the cupboard".

Having seen the parallelism between the two languages, one may presuppose that the Irish *tar éis* pattern furnishes the model for the *be after* construction in HE; however it is noted that the existence of such a parallelism does not promise identical semantic and pragmatic properties between the parallel forms in the two languages. Also, the establishment of the *be after* construction in HE will be re-examined in Where today's form came from.

Besides the view that studies on HE have offered, Heine and Kuteva (2005, p. 93–94) present the *be after* perfect as an example of "replica grammaticalization", by which they mean the case where the grammaticalization process itself is transferred from the "model" to the "replica" language. According to them, the conceptual source of the spatial/temporal schema [X is after Y] is common in HE, the replica language, and Irish, the model languages; this *after* schema is cross-linguistically rare. This may support the idea that HE speakers borrowed this construction from Irish. Heine and Kuteva suggest that it is likely that speakers of HE have adopted the Irish grammaticalization process to HE, perhaps by means of the same cognitive path that links the notion of completion with the spatial/temporal prepositional schema of *after*. In the view of the author, however, a more significant factor in the formation of the *be after* perfect is the Irish *bí* construction for expressing aspect. This is discussed in Where today's form came from.

Do be habitual

There are several examples of periphrastic *do* found in earlier texts, i.e., in John Michelburne's *Ireland Preserved*, which

dates back to 1705 (Filppula, 1999, p. 138). Bliss (1979) notes the "consuetudinal" usage of *do* and *do be* in the text, which includes (6).

(6) Why Neighour, you do be mauke de Rauvish upon de young Womans, and when…. (Bliss, 1979, p. 147)

In a later stage, in the 1860s, emigrants' letters show that the *do be* form occurs in HE at the time (Filppula, 1999). The following examples were written by an Irish mother and daughter, Nancy and Bridget Oldham, from Rossmore, Co. Cork (TCD MS 10435: *Oldham Papers*, Department of Manuscripts Trinity College Dublin).

(7) i. I do be disputing with my mother... [1857/TCD 10435-15]
     ii. I do be sick every year at this time but I was not prepared anytime until now. [1863/TCD 10435-21].

In (6), the verb *mauke* ("make") follows *do be*, while in the later stage V-*ing*/AdjP follows *do be* as in (7). McCafferty (2017) reports the following example from an emigrant letter in the 1860s.

(8) Indeed I do be thinking of ye when ye least suspect it (Schrier, 1997, p. 36; B. Colgan, 13.06.1862).

The *do be* construction in (7) and (8) is still seen in present-day SwHE.

The *'tis…* construction

The word *'tis* and clauses headed by it can be observed in Early Modern English, i.e., in the studies of William Shakespeare. The Oxford English Dictionary (OED) cites an example of *'tis* from John David (1569–1618), a renowned English poet, in which *'tis* heads a negated clause.

(9) 'Tis not that she would renovate her affection with this Prince. [1656 J. Davies tr. M. de Scudéry *Clelia* II. ii. 51]

The *'tis* in (9) is a type of *'tis* taking a *that*-clause, as in "it is not the case that…", while the HE construction is taken as a translation of the Irish copula *Is* (the "i" in Irish "*is*" is capitalized to distinguish it from *is* in English). (10) is the entry of the OED for *'tis*.

(10) ME tys, ME– tis, 15 t'is, 15–'tis (now *poetic*, *archaic*, *regional*, and *colloquial*), 16 t'is, 16 ti's, 16 tish (apparently only in representations of Irish English), 16 ty's, 19–'tes. [OED 3]

Variants of *'tis* appeared in English in the 15th and 16th centuries. *Tish* is especially noted as Irish in the OED. It is also noted that *'tis* is seen in the literature of Anglo-Irish and Irish writers including Laurence Sterne and James Joyce. Besides that, it is also common in Irish writing, including famous examples such as Frank McCourt "*'Tis*", published in 1999 by Flamingo in New York City, although *'tis* itself is not an Irish invention. (11) is an example of the *'tis…* construction from an earlier stage of development in HE.

(11) Deare Catolicks, you shee here de cause dat is after bringing you to dis place: *'tis come bourying you are de corp, de cadaver, of a verie good woman, God knows!*, fwom cruel deat hate devoure. (Bliss, 1979, p. 133, John Duntion, *Report of a Sermon* 1698).

Importantly, Bliss (1979, p. 296) notes: "Here, *'tis* corresponds to the Irish copula *Is*, the element *come burying* is emphasized, and the rest of the clause is recast as a relative clause, although the relative pronoun is omitted, as it usually is at the present day". In Bliss's collection of texts from the 1600s to 1740s, we can find examples of the variants of *'tis*, including *'tish* and *tish,* aligning with the Irish sentence-initial copula *Is*.

The Irish copula *Is* appears during the initial rise of HE in parallel contexts and settings to those where *'tis* and the variants appear. This shows us the mixture of Irish and English in the un-institutional language contact. In the *'tis* sentence in (11), the verb in infinitive form comes after *'tis*. (12) parallels this with *Is* followed by an infinitive verb.

(12) Commanded bee superior powers,
Is make me h[a]unt dese donny bowers;
And fate!, and be!, I never thought
(Bliss, 1979, p. 117; *Purgatorium Hibernicum*, 1670–75)

The Irish *Is*, usually located in the initial position, functions the same as *'tis*. (11) is an example where a verb comes just after *Is*, while in (13) *Is* is followed by the subject-NP and an adjective.

(13) …; *is none of you strong enough, or stout enough,* to overcome him, or wise enough or cunning enough, to sheet him—no, no, 'tis a shad ting, not won! Well, den, fwat's to be done?

(Bliss, 1979, p. 134–5; John Duntion, *Report of a Sermon,* 1698).

(13) uses *is* same as the Irish copula form, which would likely be replaced by *'tis* in present-day HE. It is important to note that Irish would not use the *Is* copula. (14) is another pattern in which a subject noun and a finite verb follow *Is*.

(14) I know de[e] vell enough, and bee!
Is de old hawke have de old eye
(Bliss, 1979, p. 119; *Purgatorium Hibernicum*, 1670–75)

According to Bliss (1979: 297), *Is* in earlier HE is used to mean "I am", as in *Is thanke my mester* ("I thank my master") and as in (14), and *ish*to means "it is", as in *What ish my Nation?* in William Shakespeare's Henry V in 1599 and 1623. [Bliss (1979), p. 35] notes the reference to the Earl of Essex's campaign in Ireland in 1599 and deduces the year of the text]. Bliss also mentions omission of the personal pronoun in those sentences.

In addition, *Is* takes the auxiliary *may* in the clause.

(15) Is may as velkisse my breesh
"You may as well…"
(Bliss, 1979, p. 119, 298; *Purgatorium Hibernicum*, 1670–75)

In (16), there is an example of the tensed copula, namely, *vas*, the past form meaning "was".

(16) Vas carry it on his Shoul-deer
"He carried me…"
(Bliss, 1979, p. 129, 298; *The Irish Hudibras*, 1689).

This use of *vas* is inherited from the usage of Irish *ba*, the past form of the copula *Is*. The examples (12)-(16) suggest that the *Is* pattern modeled after Irish is often used to construct sentences using English words. Over time, *Is* was replaced by the English equivalent/correspondent *'tis* or *'ish* with the information structure to be expressed held in the construction. In addition, this pattern came to be used to express information saliency, as the Irish *Is* construction does. This will be discussed in more detail in The *'tis…* construction and Expressions of information structure.

Functional properties of the forms in contemporary HE

In this section, our discussion turns to HE after it reached a certain stage of maturity. We thus highlight facets of grammar concerning the three features in focus in stages II and III in Figure 1. The following addresses the three illustrations: (i) in present-day HE, the *have* and *be after* perfects exhibit their separable functional distributions (*Be after* perfect vis-à-vis *have* perfect in HE today), (ii) the *do be* form has two discrete usages, "habitual *do be* $_1$" and "inherent property *do be* $_2$" in

the text published by a playwright of SwHE in 1950s–1970s (Two meanings of *do be* in a stabilized SwHE), and (iii) the *'tis…* construction inherits the Irish strategies of expressing informational saliency (The *'tis…* construction). The examples in this section are from linguistic interviews, from natural speech or from the John B. Keane corpus. Examples from natural speech and from John B. Keane are followed by the source in brackets. Those that do not mention the source are from linguistic interviews with speakers.

Be after perfect *vis-à-vis have* perfect in HE today

The *be after* perfect is a well-known feature of HE; however, the *have* perfect also exists in HE. The segregation in usage has resulted in a contemporary phase in which two perfects coexist. Note that the *have* perfect in HE behaves differently from StE and exhibits a much lower frequency than StE, which Kirk (2017) reveals in a comparison between London-Lund Corpus and International Corpus of English (ICE)-Grate Britain Corpus on one hand and ICE-Ireland on the other hand. In HE, the simple past tense form is often the first choice in natural speech, although the *have* perfect can be elicited for most sentences with the simple past.

A contrast between the *be after* and *have* perfects can be seen in (17) and (18) according to their different contexts.[3]

(17) {Context: Mary has cleaned her son Brian's room before he comes home from school. When he gets home, Brian throws his uniform, shoes, and school bag on the floor, and then he puts on his casual clothes and is about to go out to meet his friends. Mary sees the mess in the room. She says to Brian:}
a. # I've cleaned your room. (normal pitch)
b. Ok I'm after cleaning your room. [elicitation] (Shimada, 2010, p. 205)

(18) {Context: Ger comes home from work. He is in a hurry and leaves his working clothes and shoes on the bed. He says to his wife, Mary, *will you put away my clothes?* and goes out. He returns home again, and Mary says:}
a. Ok I've tidied your clothes.
b. # I'm after tidying your clothes. [elicitation] (Shimada, 2010, p. 205)

The *be after* example (17b) expresses the speaker's intention to act upon the listener, in addition to expressing the proposition that the activity of "cleaning your room" has been done. For example, the speaker may want to imply that "so you should

not make a mess" or "how dare you have said that and make me wash your T-shirts," or "please do not bother me so that I can relax a little"[4]. The *be after* perfect reflects a focus on the present state after the action or event is completed rather than the action or event itself. This, in turn, invites the hearer to shift their attention to the utterance-time state, and conditions the hearer to act in such a way as to address the state. Having pragmatic oppositions to the *have* perfect as such, this perlocutionary effect of the *be after* form is triggered by the grammatical form itself.

The pair of (19) and (20) is an example of a "having a guest" situation.

(19) A (host): D'you want some tea?
B (guest): *I'm after having tea.*

(20) A: D'you want some tea?
B: *I've had tea.* (Shimada, 2010, p. 187)

In (19), the speaker (guest) asks the listener (host) not to get him tea, simultaneously encoding the fact that "I have already had tea". On the other hand, in (20), the SwHE speaker's relatively neutral attitude toward the proposition of having had tea is observed.

There is another example that highlights the pragmatic difference between the *have* and *be after* perfects in SwHE. The following (21) is cited from an email text.

(21) I just started working on Monday– am working in a center for adults with learning disabilities as a psychologist – *i've just started* so idon't know exactly what they want me to do –but am enjoying the work so far! [email text, Cork City]

In (21), the italicized sentence cannot be interchanged with the *be after* perfect primarily because of the absence of the interlocutor. The *have* perfect is used for a neutral description of the situation rather than affecting the reader of the mail. (22) is the author's recent encounter of the *be after* perfect.

(22) {Context: A taxi driver picked me up on the road. She contacted the company to report this, while listening to me, and said to me after the contact:}
I was after being called, you know. [Dublin]

In (22), the utterance of *be after* was intended to encourage the listener to resume the conversation. This suggests that the "hot-news perfect" is an appropriate tag for the *be after* perfect in that the speaker of *be after* intends to share the content of his utterance as if it was a piece of news. The focus is laid on the state of the utterance time, implying something like "I am no longer talking on the phone, so please go on with what you were saying".

---

3 The examples (17)-(22) are discussed in detail in Shimada (2008, 2010).

---

4 These implications are not made by the *have* perfect. Of course, such conversational implicatures could be conveyed in an utterance containing the *have* perfect by virtue of intonation and shared background by the speaker and listener.

The preceding illustration of the pragmatic contrast between two perfects in SwHE suggests that the *be after* form and the *have* forms have developed so that speakers who use both forms assign them different pragmatic functions. This is an autonomous process in HE, apart from the approximation to the "target language" of English.

The *be after* form exhibits high productivity; it co-occurs with a wide range of verbs and nouns.[5] Its use is not restricted to any social group (Kallen, 1991). The *be after* perfect is usually labeled as "hot-news perfect."[6] In the author's observation, the label, referring to the status as news of the thing being reported (McCawley, 1971) portrays the property of *be after* V-*ing*/NP not so much because of the indication of recency that some studies (Harris, 1984; Kallen, 1989, 1990) discuss but rather because of its expressive connotations. Using *be after*, the speaker exhibits their emotional attachment to the given event completed at the time of speaking while presupposing that their addressee does not yet know the news that they are reporting (Shimada, 2010).[7]

In summary, the *have* perfect denotes completion of an action that was initiated in a prior point in time, while the *be after* construction highlights the relevance of its completion to the context of speaking. By virtue of *be*, the *be after* construction relates the fact that some activity or event is in the status of completion in the speech context, where not only the speaker but also the listener/interlocutor are involved. This difference in focus, along with the perlocutionary effect, may furthermore be supported by Kallen's survey of what type of social interaction in which *be after* tokens occur. The *be after* form is present more frequently in the conversational domain of "Friendly (work)"/"Friendly (general)"/"Family"/"Shops" than in public domains.

Two meanings of *do be* in a stabilized SwHE

The *do be* form is generally described as a habitual-aspect marker. Examples from John B. Keane's (1928–2002) literature

basically confirm this, but further examination of settings where *do be* is used has allowed for a more explicit description of the meaning and usage of the *do be* form[8].

(23) and (24) are examples of the *do be* form. The form expresses the recurring activity of the agent of the verb.

(23) 'Tis not aisy, a-girl, to kill you! You have the appearance of a small one, a young one. We *do be praying* for you in our prayers, whenever we get the notion to kneel. [SIV 21]

(24) Why *do* you *be* always *singin'* that oul song? Where did you pick it up, anyway? [MYM 1]

This usage is now called "habituality *do be* $_1$". Importantly, besides *do be* $_1$, there are a number of examples where the *do be* form appears in a predicate phrase expressing an inherent property of the subject. This "inherent property" use of *do be*, henceforth labeled as *do be* $_2$, refers to a stated quality that belongs to the object by nature. Significantly, the *do be* in SwHE occurs in relative clauses, often following an *NP* + *where*, *the way* (including *how*). This kind of clause, appended to a head NP, defines an inherent property, offering a complementation of the antecedent.

(25) Will you open it or you'll drive me to *Gleannnan Gealt* where your own equals *do be*. [SIV 39]
   "*Gleannnan Gealt* where your own equals are (*Gleannnan Gealt* is known to be the place where your own equals are)"

(26) 'Tis a wonder you took your backside from *the table* where people *do be eating*. [SIV 3]

(27) What would be in it but thoughts to disturb her young head the night before her marriage. Have you no knowledge of the way *a woman do be* the night before? [SIV 38]

In (25), the place name *Gleannnan Gealt* is specified in the relative clause; similarly, in (26), the inherent property of "the

---

5   As to the wide range of verbs, Kallen (1991) reports that his Dublin corpus, which contains 114 tokens of *after* obtained from 74 speakers in Dublin, includes 56 verbs, with the following frequencies: *being* (10 tokens of copula and seven tokens of auxiliary/passive) and *getting* (12 tokens).

6   This form has been labeled as "hot news perfect" Harris (1984), the "*after* perfect" (Filppula, 1999), and "immediate perfective" (Hickey, 2000).

7   This is based on the author's sessions with speakers in years 2004-2007. In the impression the author has formed from visiting Ireland regularly, the usage of *be after V-ing* has expanded and generalized to include a use in non-expressive contexts (Shimada, 2016b, p. 163–164). Ronan (2005) notes that, the range of use of the *after* perfect is speaker-specific.

8   Keane is a playwright from Listowel, County Kerry, where the author has been undertaking fieldwork since 2003. Examples from John B. Keane's plays and letter series written from the 1950s to 1980s are cited with their abbreviations; the collection of examples is referred to as the "Keane corpus". The examples are sorted by grammatical features and categories; one of the features is' *Tis(~it is)...* constructions. The Keane corpus, as a whole, comprises over 18,000 words. Keane is a local writer and his work, cherished by local people, is an invaluable source of the local language from the 1950s-1970s. He is known as a major Irish writer with many successful plays and books (Smith and Hickey, 2002). The following is a list of his woks cited in this article, headed with their abbreviations: SIV, *Sive* (Keane, 1959); HHM (Keane, 1961); STD, *Letters of a Successful TD* (Keane, 1967). The spellings used in the citations are those used in Keane's plays. This article also includes examples that the author happened to record and encounter during his fieldwork.

table" is specified in the clause. (27) is an example of *do be₂* occurring in a clause followed by the antecedent "the way".

The *do be* habitual is considered to have developed elaborate usages until it constitutes a part of the stable system of SwHE. Shimada (2013) reports that while there are a number of examples of *do be₂* found in Keane's play scripts in 1950–1970s, this usage is not attested by present SwHE speakers. Uses of the *do be₂* type have largely converged with those of the unmarked present tense in contemporary HE, and the convergence, combined with the speakers' awareness toward standard norms, may lead to a relative decline in usage; this will be further addressed in Norm shift by the second contact. The presence of two types of *do be* in the Keane corpus and the probable decline in stage III may foreshadow its change and relocations between form and function in the system of SwHE.

The *'tis...* construction

(28) and (29) are examples of SwHE cited from Filppula (1999), whose informants were born sometime in 1900–1910, with the recordings conducted in 1970s. (30) is an example from the same county (Co. Kerry) of the author's Keane corpus.

(28) a: And did they speak English and Irish?
      I: There's more spoke Irish one time. It have died away. Our language is dying away, since we got our own, independence. *It is more English they are speaking now.* [Kerry] (Filppula, 1986, p. 93)

(29) …and I here in bed with my nerves in a bad state and my left breast sore. Maybe *'tis cancer I'm getting* or maybe 'tis something wrong altogether. [STD 56]

(30) How do we know but maybe 'tis dead you are, or worse. [STD 11]

These types of sentences are often described as *it*-clefts in the literature because of similarities in appearance to the *it*-cleft in British English. However, because of some stark differences with BrE *it*-clefts, the author gives them their own label. Since, in SwHE, a majority of the examples have procliticized forms beginning with *'tis* (e.g., *'tis, 'twas, 'tisn't*) and separable properties from *it*-clefts, the author refers to them as the *'tis…* construction.

Studies have noted that in HE, "clefts," the *'tis…* construction in the author's terminology, occur highly frequently (Curme, 1931; Taniguchi, 1972; Filppula, 1999) and allow for wider ranges of phrasal categories in the focus position than they do in British English (e.g., Guilfoyle, 1985; Filppula, 1986, 1999; Henry, 1995). For example, the phrasal categories AdjP and non-finite VP can occur in the focus position in HE. Besides the phrasal categories in the focus position, there are important differences in form and function of the *'tis…*

construction. Examining the characteristics of HE and how they contrast with StE will allow for us to begin to see the sentence construction and the way of expressing information structure in HE.

(i) *'Tis*

'Tis is prominent in SwHE. *'Tis* (pronounced as /tiz/) is a procliticized form whose major tensed/negated representations are *'tis, 'twas,* and *'tisn't*. It is recognized by SwHE speakers as a single unit. *'Tis* is frequently used in SwHE daily conversation, both in cleft-like patterns and apart from them. In particular, *'tis* forms are often used as answers to yes-no and tag questions, and as affirmative responses to sections of dialogues. (31)–(32) illustrate these usages.

(31) A:*'Tis* a grand day thank God!
      B: *'Tis! (with a nod)* [Cork]

(32) Ellen: But Glory be to God is anyone safe? That's a dead loss.
      Nora: *Tis.*
      Ellen: That's a dead loss that place.
      Nora: *Tis tis.* [cited from (Murphy, 2006)]

This usage of *'tis* as a single-word utterance may gain the idiomatical status of a response token. It is common among sentences containing the sentence-initial *'tis* in HE. In HE discourse for giving responses, *'tis* is often used as observed in (33), cited from the International Corpus of English (ICE)-Ireland (Kallen and Kirk, 2008).

(33) B: A lot of, a lot of them now we get, come from Beechwood or, some of these […] yeah yeah. And they can't understand, that the measure of care they get, by comparison with paying a wad of money.
      A: Mm
      B: For nothing.*'Tis only... basically... get them up and feed them* [ICE-Republic of Ireland, S1A055]

(33) is a conspicuous example that reveals continuity of the function of *'tis* (*'twas*) between its use as an independent lexeme, seen in (31) and (32), and its usage in the *'tis…* construction. Both usages appear in (34).

(34) I: But *'twas the tenants put up that*, like, his tenants. I believe, that put up that.
      b: I didn't know that.
      I: *'Twas, 'twas*, I don't know, but it is written on it, that *it is his tenants put up that*, like. (Filppula, 1986, p. 171)

The first sentence, in which *'twas* is used, is likely to have been spoken to connect things in mind with the preceding utterance. In the third line, *'twas* is used to recall things in mind. The following *it is* may reveal the speaker's affirmation that the event (i.e., his tenants put up that) is true.

Furthermore, the continuity of the discourse function of the *'tis…* construction can be observed in the usage of *'tis* in (35), which the author happened to hear in a conversation during his fieldwork in Kerry. B is explaining why A should read the book.

(35) A: Why do you say so?
      B: *'Tis everything in that book happened in this area.*

The *'tis…* construction is used to give a response to the question in (35). The responding speaker indicates her strong commitment to the truth of the thought.

The *'tis* in HE can be described as a discourse marker for affirmation, showing subjective commitment to the statement. The word "subjective" is used here as an adjective derivation of subjectivity, which means the "expression of self and the representation of a speaker's perspective or point of view in discourse" (Finegan, 1995, p. 1), in other words, a "speaker's imprint" (ibid.). This article, for the current purpose, does not pursue the issue of subjectivity any further; however, it suggests that *'tis* is used as a marker of the subjective commitment to the content to be expressed by the utterance that the said marker introduces.

(ii) "Salience sensitivity"

This article maintains a "non-cleft" analysis suggested by Shimada (2018) for the *'tis…* construction in present-day SwHE. The "non-cleft" means that the analysis does not depend on *it*-clefts in StE. The description was made not on the norm of StE but on the norm of SwHE on its own to allow for a more economical description of SwHE. This permits a uniform explanation for the HE data on its own. "[T]he initial assumption is that if a linguist who has never spoken or learned any variety of English encounters HE for the first time for linguistic description, (s)he could find a better, or at least different, way of describing *'tis(~it is)…* and the things involved in light of the data from this language (Shimada, 2018, p.249)." According to Shimada (ibid.), the *'tis…* pattern can be accounted for as the combination of the clause-initial *'tis* and a finite clause, in which salience marking may operate. She defines "salience" as having a heavier informational load than other elements in the state of affairs that is being expressed in a sentence. Salience can be marked in speech syntactically and prosodically. HE is salience-sensitive; informational saliency determines not only prosodic presentations but also syntactic forms.

To make the concept of salience more explicit, Shimada (2018) compares salience and focus using pitch accent and the *it*-cleft in StE. In general English, prosodic prominence is used to mark saliency of the information on a particular constituent in a sentence. For example, a speaker may say "I saw MARY in the park" with prosodic prominence on *Mary* when they place more significance on the Mary constituent than the other syntagmatically related constituents. That is, *Mary*

is salient. On the other hand, the syntactic strategy of using an *it*-cleft, as in "it is Mary that I saw in the park", expresses focus. Focus, more precisely identificational focus, is defined as a new piece of information *vis-à-vis* the presupposition made by the sentence, while salience is, by definition, independent from the presupposition and addresses the syntagmatic relation of constituents. There is, of course, overlap between focus and salience, but salience has proven to be a more economical way to precisely describe the function(s) of the *'tis…* construction in SwHE. This "salience sensitivity" comes from the Irish language, as will be further considered in Expressions of information structure.

While the *it*-cleft is a syntactic expression of focus, the HE pattern that resembles *it*-clefts on the surface is the combined outcome of the *'tis(~it is)* clause and the fronting of a salient constituent. HE is salience-sensitive; informational saliency determines not only prosodic presentations but also syntactic forms.

(iii) The sentence construction

'Tis… clauses the following: giving the reason or adding to the topic in the previous utterance, i.e., (36); assuring one's self or someone else, (37); recollecting an event in the past or opening a discourse, (38).

(36) I do like to be beside the seaside. *'Tis far away from the seaside we are*, God bless us and save us in the warm weather. Far away indeed from the time I was a slip of a girl walking the streets of Ballybunion in my figure and you winking at me. [STD 56]

(37) …and I here in bed with my nerves in a bad state and my left breast sore. Maybe *'tis cancer I'm getting* or maybe *'tis something wrong altogether.* [= (29)]

(38) (In the beginning of a column) *'Twas at the river bank I met {X=name}.* She stopped to talk to my dog. I told her it wasn't my dog and that I was walking Dougal for a friend. [Cork]

'Tis takes a clause in which a salient constituent, in syntagmatic terms (i.e., compared to other constituents of the sentence), is fronted, if there is a one as such.

In HE, salience is marked by fronting a salient constituent in cooperation with prosodic prominence, sometimes with the supplement of *that* placed after the salient constituent of NP-subj, PP, and AdvP. In addition, the insertion of *that* significantly lessens acceptability and the context that the construction appears different from that of the *it*-cleft in StE as discussed.

(39) a. 'Tis dead you are.
      b. *'Tis dead *that* you are.

(40) a. 'Tis grumbling he is.
      b. *'Tis grumbling *that* he is.

The use of *that* in the case of AdjP and non-finite VP salience is ungrammatical or infelicitous, as noted in (39) and (40). A data survey that the author conducted confirmed that *that* is present when the chunk of the fronted constituent (NP-subj, AdvP, and PP) is large relative to the other constituents, as in (41).[9]

(41) He's doing well for himself when you consider he left the national school from the fourth class. *'Twas from studying the television programmes in the papers that he learned how to read.* [STD 10]

The boundary of the salient constituent *from studying the television programmes in the papers* is clearly expressed with the support of *that*. Another notable feature is that *that* occurring in this pattern is spoken with or without a break after the salient constituent on some occasions (e.g., in reading a text aloud), and it is also true that SwHE speakers find the *that*-less sentence *'Twas from studying the television programmes in the papers he learned how to read* preferable in spoken language.[10] *That* occurs when informational saliency is not entirely expressed with the fronting of a salient constituent in the *'tis* complement and prosodic prominence of the fronted constituent. It is used to supplement salience marking.

Irishness realized in English

Expressions of tense/aspect

The *be after* perfect is generally supposed to have calqued on *tar éis* (∼*tréis*) in Irish. Although the *be after* construction seems to be a straightforward translation of the *tar éis* construction in Irish, the low frequency of *tar éis* in Irish (Greene, 1979; Ó Sé, 2004) makes it a poor candidate for calquing. Greene (1979) reports that there was only 1 example of *tréis* in 19,753 verbal forms in *Buntús Gaelige* of the 100,000-word corpus, in addition to his reference to Hartmann (1974) who reveals that there was no example in a corpus of conversations totalling about 117,000 words recorded in Connemara. Ó Sé (2004) notes that there are only four *tréis* examples of the 360 of verbal adjective constructions in *AntOileánach*.

Greene (1979) also reports that use of the construction using *tréis* as in the example (42) is rapidly extending its field in Irish spoken outside the Gaeltacht. This suggests influence on Irish from HE.

(42) Tá sé tréis leitir a scríobh.
      "He's after writing a letter."[HE] (Greene, 1979, p. 122)

Tréis used in a substantive verb clause had not always been an idiomatic or grammaticalized way of expressing aspect, evinced by its low frequency in corpora. The sudden increase in the use of *tréis* outside the Gaeltacht, in addition to the evidence of low frequency in corpora, implies that the expression of perfect using *tréis* in Irish is gaining in use now, so it is more of a modern expression. This, by extension, gives support to the idea that the *be after* perfect is an invention of HE on its own. If not an idiomatic use of *tréis* in Irish, what was involved in forming the *be after* construction? The author located the answer in the Irish method of expressing aspect. Irish has verb-initial construction, and this is likely to have provided the schema for HE's sentence production. Rather than being a calque from *tréis*, it is more likely that the general characteristics of sentence construction in Irish have influenced the formation of HE. In other words, as the author will argue below, the *be after* construction is construed in the stability of the subjunctive verb construction in Irish, which allows for aspectual expressions.

In Irish, aspect is denoted by constructions with the substantive verb *bí* (*tá* in the present tense) in the sentence-initial position, one of which is the passive construction (Stenson, 1981; Ó Siadhail, 1989; Russell, 1995; Doyle, 2001; Nolan, 2006). Note that the passive used here is particular to languages with substantive verbs and refers to the sense where "a noun phrase, which does not present the agent, appears as the subject of the substantive verb in the first argument slot following the substantive verb in the position reserved for the grammatical subject" (Nolan, 2006, p. 140). The term "passive" is used in the literature (Ó Siadhail, 1989; McCloskey, 1996; Ó Sé, 2004) in reference to many European languages as a syntactic term concerning the realization of arguments; in the passive in Irish, the argument corresponding to the subject of an active verb appears as the object of the preposition *ag* "at" (McCloskey, 1996).[11] "Perfect" has been generally adopted by grammarians of Irish (Ó Siadhail, 1989; Ó Sé, 2004; Stenson, 2020). There are

---

9   The author's data looked at the constituent in the position following *'tis*, represented as X in the following. In the author's data, there are respectively in [X (supposed focus position) = NP-Subj] 48 examples of the *that*-less type and 60 tokens with *that*; in [X = PP/AdvP] 60 and 21, in [X = NP-Obj] 43 and 1; in [X = AdjP/Non-finite VP] 18 and 0. Data presented in other studies, although they do not give numerical information, also turned out to this distributional tendency (Filppula, 1986, 1999; Ó hÚrdail, 1997).

10   Some SwHE speakers who pronounce *that* rather clearly in AdvP-saliency sentences report that such examples are "modern" [Co. Kerry]. This may suggest the influence of their awareness of normative grammar. It is assumed that this awareness motivates the use of *that* in HE.

---

11   The passive status of this construction may be questioned in terms of how it is formed and the fit of the passive in the system of Irish. Dillon (1941), claiming that Modern Irish *aa sa daantaagam* ("I have done it") is not passive, points out that the substantive verb was construed with the past participle already in the Old Irish period. Dillon further argues that "the participle is an adjective and the verb "to be" is the main verb, not an auxiliary" (1941, p. 59). In contrast, McCloskey (1996, p. 254–255) argues that "in a full clause, the participial form is supported by the verb *be* used as an auxiliary; in a small clause, the participle appears without any auxiliary". Ó Sé (2004), to whom this article adheres, basically agrees

some dialectal variations over the choice of active and passive variants. "The passive variant has a far higher frequency in Munster" (Russell, 1995, p. 102), the area where SwHE is spoken. In Munster, the aspects of the progressive, the prospective, and the perfect are expressed in the passive construction. (43a, b) and (44) are cited from Nolan (2006, p. 141–142), and (43 c, d) from Ó Sé (2004, p. 181); the glosses are the author's. English translations are given based on Greene (1979), Ó Siadhail (1989) Russell (1995), Doyle (2001), Nolan (2006), and Stenson (2020). (43) shows perfect forms, (44) prospective forms, and (45) progressive forms.

(43) a. Tá an leabhar léite agam. [perfect, Irish][12]
       be.pres. *the book read.*$_{VA}$*at-me*
       "The book is read by me./ I have the book read [HE]".
       b. Tá an leabhar léite.
       be.pres. *the book read.*$_{VA}$
       "The book is read".
       c. Tá sé tagtha
       be.pres. *he come.*$_{VA}$
       "He has arrived".
       d. Tá sé feicthe ag Máire.
       be.pres. *it seen.*$_{VA}$ *at Máire*
       "Mary has seen it".

Ó Sé (2004, p.186) has left a note that "evident increase in frequency [of the perfect constructions, especially the past participial types in southern Irish, as in (43 c, d) of our examples] in the twentieth century is doubtless due to bilingualism with English". Ó Sé (2004, p. 186) also informs us of the fact that "until quite recently, the perfect was considerably less frequent in Gaelic dialects than in English or other languages of western Europe". Given these insights, we shall extend our view to other aspectual constructions in Irish that are similar in the use of the peripheral *ag* (=at) + agent (or actor).

----

with McCloskey (1996) except that he adopts "perfect passive" instead of "perfective passive" to describe this construction.

12 This aspect has been referred to in the literature with various terminologies such as "perfect," "perfective", and "completive" (Ó Siadhail, 1989; Ó Dochartaigh, 1992; Mac Eoin, 1993; Russell, 1995; McCloskey, 1996; Ó Sé, 2004; Nolan, 2006; Stenson, 2020). In light of the fact that the state-of-affairs that the verb phrase expresses is a completed state with current relevance, the author chooses perfect, not perfective. His use of "perfect" and "perfective" is based on Comrie (1976, p. 12): "perfect" refers to a past situation that has present relevance, for instance, the present result of a past event; whereas "perfective", contrasting with "imperfective", denotes a situation viewed in its entirety without regard to internal temporal constituency. Following Comrie, to avoid ambiguity between the two meanings, the author avoided using "perfective" as the adjectival form of "perfect". Note that while "perfect" focuses on the state resulting from the completion of an action, the "completive" focuses on the completion itself, and so it can be applied to, i.e., (43).

(44) a. Tá an leabhar leléamh agam. [prospective, Irish]
       be.pres. *the book with reading.*$_{VN}$*at-me*
       "The book is to be read by me".
       b. Tá mé le/chun léamh an leabhar.[13]
       be.pres *I with/toward reading.*$_{VN}$*the book*
       "I am to read the book".

(45) Tá an leabhar (dh)á léamh agam. [progressive, Irish]
       be.pres *the book at reading at-me*
       "The book is being read by me/ I'm reading the book".

These periphrastic sentence constructions seem to have provided Irish speakers who were shifting to HE with the sentence-construction schema for expressing aspectual meanings. Considering the deep-rooted use of the *bí* construction and its multiplicity of uses in Irish, a plausible scenario is that it provided a schema in the incipient days of HE for creating an additional aspect using "after" in the expansion of the pattern [*bí* (= substantive verb) NP$_{undergoer}$ PP *ag* NP$_{actor}$]. Also noted is that aspectual categories that are expressed in the *bí* construction are transferred from Irish to HE. Thus, in HE, Irish semantic divisions are realized using the forms provided by English, conforming to the English morphosyntax in an SVO language where the agent becomes the subject of the sentence.

Since the form of *tar éis* is not used as an aspect grammatical marker in Irish, the *bí* constructions, which served as the core of aspect-marking in Irish, are the most likely candidates for producing HE-specific expressions for aspect. Specifically, *bí* constructions inspired the innovated use of *be after* in HE to decode a completive state after the action that the verb expresses at the time of utterance. Although the use of passive constructions for denoting perfect, prospective, and progressive aspects in Irish could not transfer directly to HE where the agent or actor who does the action that the verb expresses usually becomes the subject of the sentence, it has had some influence by negotiation with the expression frame of English.

Explaining the habitual *do be* of HE also requires reference to the Irish language. In HE, variations are seen in realization of the habitual aspect, with, roughly speaking, *be(es)* in the north and *do be* in the south. The existence of the habitual forms in the different dialects suggests that the habitual category in Irish paved the way for its formation. There are numerous descriptions concerning the origins of this feature in HE (Joyce, 1910; van Hamel, 1912; Bliss, 1972; Harris, 1986; Kallen, 1986; Filppula, 1999; Hickey, 2000). The substratal presence of the category created a need by the speakers for its linguistic expression. This conditioned their reanalysis of the periphrastic *do* in primary linguistic data of Early Modern English as marking habitual aspect. In addition, Kallen (1985) has postulated that

----

13 There are dialectal variations for the use of the preposition *le*. In Munster, "there is a distinction between obligation using *le* and intention using *chu(i)n* 'toward'" (Russell, 1995, p. 100).

the *do be* form results from re-interpretation of *do* of Early Modern English, which was then juxtaposed with *be* in HE to mark a habitual, durative, or generic aspect. A similar analysis is given later by Hickey (2000) under the term "usurpation" (ibid: 113). According to him, periphrastic *do* forms that were semantically empty in the input variety of English were functionalized in HE because of a habitual grammatical category in the Irish substratum.

In Modern Irish, habituality is expressed as in *bíonn*, the inflectional form of *bí*. The present habitual *bíonn* indicates a repeated or regular state of affairs.

(46) Bíonn sé tinn.
      "He is (regularly) sick; he is sickly" (Stenson, 2020: 59).

The existence of a morphological category for the habitual in the form of the substantive verb would have been important, as it would inspire the need for the equivalent category in the newly adopted language. Also, Bliss (1972, 1979) argues that Irish speakers acquiring English associated periphrastic *do* with the dependent form ending in Early Modern Irish *-(e)ann* (present indicative) on the basis of their contextual parallels; the use of this ending, which originally functioned as a dependent form, was extended into the absolute environments over the course of the 17th century. "[T]he ending *-(e)ann* is found only in the consuetudinal present *bídheann*"; the auxiliary *do* "would therefore come to be associated with a consuetudinal meaning" (Bliss, 1979, p. 293). Harris (1986) criticizes this hypothesis, pointing out discrepancies between syntactic contexts in which the Irish-dependent form and StE periphrastic *do* occurs and explains that the dependent form in Modern Irish occurs not only after the particles *ni* (negative), *an* (interrogative), and *nach* (negative-interrogative), as the expanded use in the contexts where *do*-support is required but also in other contexts such as after the conjunctions *mura* ("unless"), *dá* ("if"), *go/gur* ("that"), and the relative particle *a*. The characteristic he points out actually shows a striking parallel with *do be₂* of SwHE from the Keane corpus, which only appears in relative clauses, as we observed in the section *Do be* habitual.

Concerning the use of *do* in English in Ireland, the periphrastic *do* remained in the nonstandard daily spoken English vernaculars in south Ireland, whose use of *do* was strengthened under the linguistic conditions of Gaelic/English bilingual speakers[14] despite the assumed decline until the 18th century in south England. According to Filppula (1999, p. 140–142), based on a thorough survey of studies on the use of periphrastic *do* in BrE, "this construction [periphrastic] reached its peak in the middle of the 16th century, after which it started

_____

14 The author's assertion here is bolstered by Harris' insights. He notes (Harris, 1986, p. 193): "it can be argued that the distinctive habitual markers in both the substrate and nonstandard varieties of the superstrate had a mutually reinforcing effect on the development of a similar category in the new contact vernaculars".

to decline very rapidly and became quite rare by the early 1600s" (Ellgård, 1953; Filppula, 1999, p. 140). Rissanen (1991), in a study based on the Helsinki Corpus, "notes the significant drop in frequencies of use of periphrastic *do* in the records of trials (which can be considered to be closest to the spoken mode) as early as the period 1570–1640, whereas the pattern retained a relatively high frequency of use in official letters even in the last EModE [Early Modern English] period, i.e., between 1640 and 1710" (Filppula, 1999, p. 141). There is also a note in Wright (1900) that periphrastic *do* "became obsolete after about 1700 (apart from archaic and poetic uses), except in the south-western dialects where it survives as the normal form up to the present day".

In Irish grammar, the opposition between permanency and temporality is primarily significant; this is most evident in two types of verbs for "to be", namely, copula and substantive verbs. (47) illustrates the fundamental distinction.

(47) a. Is bainisteoir mé.
      "I am a manager".
      b. Tá mé i mo bhainsteoir.
      "I am a manager (lit. I am in my manager.)" (Stenson, 2020, p. 51).

In (47b), a construction consisting of the substantive verb *bí* (*tá* in the present form) and a prepositional phrase are used to indicate the temporal status of being a manager. Most relevantly is that the substantive verb is used when temporal reference other than to the speech time is needed. This property of *bí* provides the basic outline for the expression of various kinds of aspect.

Aspectual distinctions such as perfect, progressive, prospective, and conditional are subsequently and periphrastically expressed by a combination of the substantive verb *bí* and a preposition with a verbal noun phrase (Stenson, 1981; Ó Siadhail, 1989; Russell, 1995). It is noteworthy that habituality is, on the other hand, morphologically expressed with the inflected form of *bí* (i.e., *bíonn* in Modern Irish). In addition to analyzing the fit of *do be* in the overall aspectual system of HE, which is the perspective taken by this article, a specific look at the historical formation of the *do be* form in HE provides insight into the rationality of the grammar system of HE. Importantly, English did not use the auxiliary *do* with *be*, and there was no parallel in Irish for the HE *do + V* pattern (Filppula, 1999, p. 137). Filppula (1999) discusses the pattern involving *be* in the *do be* form, referring to Bliss (1972) argument: Phonetic resemblance between Irish *bí/bídh* and English *be* would facilitate transfer of the use from Irish *bí/bídh* to English *be* by Irish learners of English and thus provide the basis for the adoption of *be* as a consuetudinal aspect marker in HE. The dependent form in Irish had a syntactic distribution very similar to the uses of the auxiliary *do* in English, i.e., in early Modern Irish "be", the ending *(e)ann* was only found in the consuetudinal present *bídheann*.

Furthermore, by taking into account the oppositions in the Irish verbal system, we could seek the answer to the question "why *do be V-ing*, not *do + V*, was innovated and has survived in SwHE? The difference between *tá* and *bíonn* is most likely to be displayed in the semantic distinction between "I am writing a letter" (but not "I write a letter.") and "I do be writing a letter" in SwHE. Owing to the substantive construction expressing both continuous (*bí*) and habitual (*bíonn*), the *do be V-ing* form, having *be* within, is more stable than *do + V* when expressing habituality and has thus become an established part in contemporary varieties of SwHE.

Expressions of information structure

We are now in a position to consider why the *'tis....* construction has the form it does today. As discussed in the sections of *'Tis* and "Salience sensitivity", it holds distinct properties from the *it*-cleft of English, which it resembles on the surface. We examine Irish traits and their integration into forms of English *via* the *'tis...* construction in SwHE. The claim that HE has inherited its way of expressing information structure from Irish. In other words, in both languages, salience affects the form of sentences. Recall that salience refers to the syntagmatic relationship between constituents. In a salience-sensitive language, a constituent having more saliency, in theory, is marked prosodically and/or syntactically. The following gives evidence that the salience dependency of SwHE is traced to the Irish language.

We have adopted "salience" rather than "focus" to describe the information structure of *'tis....* In describing salience, the expression of salience in StE *via* prosodic accent and the tendency of HE to express this syntactically were discussed. This discussion is now extended to the Irish language. Greene (1966, p. 42) states that "Irish expresses emphasis by grammatical means rather than by intonation, and any stressed word can be brought to the head of a sentence, with *is* [*Is*] before it"[15]. Prosodic marking is not a general strategy for denoting information structure in Irish, but it adopts the syntactic device using *Is*.

_____

15   Likewise, Cotter (1994, p. 134), contrasting the Irish and English systems, notes that, "Much of what is signaled by intonation in English is encoded grammatically in Irish. In particular, the most "important" aspect of an utterance will be moved toward the front of a sentence in Irish or be marked morphologically". Cotter, for the purpose of illustrating the "focus-marking" system in Irish, defines focus as a "highlighting or emphasis of a particular constituent in an utterance" but "not in relation to presuppositions or background". This means that in order to illustrate the Irish "focusing" constructions, the primary definition of focus must be discarded. In the description of information structuring in HE, a similar change or adjustment compromise is necessary. Rather than redefining focus, this article addresses salience.

Let us have a closer look at examples of HE and Irish. In HE, salience can be syntactically expressed even when not accompanied by *'tis*. (48) is an example of fronting in SwHE.

(48)   Pats: I saw the young girl, Sive, and the other one going the road to town airly [early] in the day.
       Nanna: *Gone to buy the wedding clothes they are. Fifty pounds Dota gave to buy the clothes and the drink for the wedding.*
       Pats: 'Tis about the wedding I came. Last night we made a plan in the caravan. [SIV 34]

Marked constituent order, underlined in the conversation above, is often used for expressing informational saliency in SwHE,[16] and this strategy is one of the elements manifested in the *'tis...* pattern. In StE, on the other hand, this type of informational saliency is likely to be prosodically encoded.

The analysis of the *'tis...* pattern as a combination of *'tis* and a clause in which salience marking may operate is similar to the behavior of the corresponding Irish constructions. *'Tis* functions as a discourse marker to mark the commitment of the speakers' subjectivity. Omission of *'tis* does not affect grammaticality in case of a *that*-less construction. (49a) resembles the *'tis...* pattern in HE; (49b) shows fronting of the salient constituent without a copula.

(49)   a. *Is* lúchorpán a chuartaíonn Seán. [Irish]
          COP leprechaun REL seeks John
          b. Lúchorpán a chuartaíonn Seán.
          leprechaun REL seeks John (McCloskey, 1979, p. 116)

In Irish, the clause initial *Is*, a copula, can be omitted, as in (49a). The translation is given in general English by McCloskey as "It's a leprechaun that John seeks" for both (49a) and (49b). The information structures are, however, more apparent in their HE translations in (50).

(50)   a. *Is* lúchorpán a chuartaíonn Seán. [Irish]
          "Tis a leprechaun John seeks" [HE]
          b. Lúchorpán a chuartaíonn Seán. [Irish]
          "A leprechaun John seeks" [HE]

The Irish *Is* construction is straightforward to be translated into HE, unlike StE, thanks to the similarity between HE and Irish in the syntactic expression of salience. Further examples are shown in (51) and (52); the translations of which the author has provided in HE and StE.

_____

16   The basic word order of HE, including SwHE, is SVO. It is often said, however, that the order of HE is not so rigid because "topicalisation is more frequently used in the rural varieties of HE" (Filppula, 1990:44). "Topicalisation", in Filppula's words, refers to the movement of a constituent from an unmarked to a marked, i.e., the clause initial, position. According to his data on the frequencies of topicalisation per 1,000 words (1986: 190-94), Kerry and Clare, where SwHE is spoken, show the highest level (1.4) compared to HE dialects and British English (BrE) [cf. Dublin 0.9, BrE 0.4].

(51) Is ag caint a bhíodar. [Irish]

Is at talk.$_{VN}$ PRT *were.they* (Doyle, 2001, p. 89)

"Tis talking they were" [HE].

"They were TALKING./(*It is talking that they are.)" [StE].

(52) Is é an t-arbhar a bhaineann m'athair le speal [Irish].

Is it the grain REL *reaps my father with scythe* (Cotter, 1994, p. 136, 139)

"Tis the grain my father reaps with a scythe" [HE].

"My father reaps the GRAIN with a scythe/(It is the grain that my father reaps with a scythe)" [StE].

The correspondence between *'tis* and *Is* is supported by an independent observation of Greene (1966, p. 40–42) concerning Irish. He observes that Irish, from its oldest form[17], always uses *Is* in a way which resembles *it's* and *c'est,* as opposed to *is* and *est*, in modern English and French, respectively. Greene (1966, p. 40) states: "Even in the oldest Irish it is already used in much the same way in which *it's* and *c'est* are used in modern English and French; these constructions, like *Is*, represent a departure from the old rule that the subject and predicate of the verb 'to be' must be in the same person, as we can see clearly from the sentence *It is I* in the English Bible (Matt. xiv 27), where the Greek original has *ego eimi* and the Vulgate *ego sum*".

In other words, *Is* is used impersonally. The literal English translations of Greene's examples in (53) below capture this property of the Irish *Is*, where he gives *it's* for the translation of *Is.*

(53) a. Is múinteoir é. [Irish]

"It's a teacher him/He is a teacher".

b. Is é Seán an múinteoir [Irish].

"It's him, Seán, the teacher/Seán is the teacher".

Greene also points out that in the identification sentence (54a), the words *Is múinteoir* were translated as "it's a teacher", and they can also be used in emphatic sentences including (54b).

(54) a. Is múinteoir Seán.

"It's a teacher, Seán".

b. Is múinteoir atá ina chónaí anso.

"It's a teacher who lives there".

The *Is* used in identification sentences such as (53a,b) and (54a) is likely to be the same as the *Is* used in the emphatic sentence in (54b), which would mean that *Is* is always used in an auxiliary role in salience-marked sentences. The shared impersonalness of *Is* in Irish and *it is* in English allowed for HE to develop its own uses of *'tis* and its variants *it's* and *it is* for the salience-aiding function.

In this way, the *'tis* of HE and the *Is* of Irish show significant similarity in the expression of information structure. In addition, they allow a similar range of phrasal categories

---

17   Thurneysen (1946, p. 327), a grammarian of Old Irish, also describes *Is* as "it is", corroborating the similarity.

in the position of a salient constituent. As in HE, in the Irish pattern (*Is* + X + relative particle *a* + Y), NP-Subj, NP-Obj, AdjP, PP, and AdvP can occur in the X position (McCloskey, 1979, 2005; Stenson, 1981; Ó Siadhail, 1989; Doyle, 2001). It is also noteworthy that the sentence-initial usage of *'twas, 'tisn't, 'twasn't, isn't it,* and *is it,* respectively corresponds to the Irish copula forms of past *ba,* non-past negative *ní,* past negative *níor,* negative interrogative *nach*, and the interrogative *an.* The Irish copula construction was transferred to HE and has developed into the *'tis* construction used to mark speaker's commitment to the proposition the clause expresses. As a whole, *'tis* marks affirmation, *'twas* recollection and *'tisn't* negative confirmation and so on. Having the copula-derived marker in the sentence-initial position, HE has inherited the strategy of mapping information structure onto sentence forms from Irish and realized it using English morphosyntax.

Contact-induced grammar formation

HE has been referred to as a "contact vernacular", and the resemblance of its historical background to that of creoles has been discussed by some linguists of HE (Harris, 1984, 1990; Filppula, 1990, 1999; Odlin, 1997; Todd, 1999). Harris (1991, p. 314), for example, states that, "HE is not typical of English vernaculars in that it has a recent history of language contact which at least partially resembles those of creoles". Although the majority of writers on HE "fight shy of labeling HE as a creole" (Filppula, 1999, p. 15) and thus designating it as a distinct language, it is also a fact that HE cannot be fully understood if it is described only as a local dialect of English. In the initial development of HE, it becomes necessary to regard HE as an independent product resulting from the synthesis of English and Irish. Moreover, the claim that HE is not a mere dialect of English is sustained by empirical evidence in linguistic and sociolinguistic data of HE. Many of the characteristics would be difficult to explain without the influence of an Irish substratum, as already argued in pioneering studies, especially Filppula (1999) among others. The attempt in this section is to visualize the process of formation of HE as a contact vernacular, a process that continues to the present day as discussed in prior sections.

Studies of other scholars from contact-linguistic perspectives also support this analysis of HE, leading us to consider the link between so-called creole and non-creole languages like HE and focus on language phenomena. DeGraff (2005), i.e., disaffirms the distinction between creole and non-creole languages. Similarly, Winford (2001, 2009) discusses basic similarities between English-based creoles and indigenised varieties of English in the characteristics of their paths of development, both of which, he claims, were shaped by the interaction of L1 knowledge and universals of language creation. Inspired by these discussions, this article proposes a model of contact-induced grammar formation

based on HE, which has obvious involvement of the two source languages, Irish and English. The following model is suggested based on the analyses presented in this article, which were greatly helped by the findings and analyses of HE in major studies by Harris (1984) and Filppula (1999). The approach of the author hereafter highlights changing linguistic aspects of HE as they relate to the likewise changing socio-historic situation, and makes no commitment as to whether HE should be deemed a creole or a dialect.

Where today's form came from

So far, we have observed some earlier examples in the rise of HE and some of the contemporary facets of the system behind it. We have discussed Irish language traits and their realization in HE morphosyntax. A common factor among *do be, be after*, and *'tis…* is that they reflect grammatical categories inherited from Irish. Irish offers a base for grammatical oppositions, which can be described as conceptual constructions dividing the world up in order to express it with language. This base is involved in the formation of a sentence by forming the basis for grammatical categories, with which states of affairs are rendered.

For example, we have seen that habituality is grammaticalized both in Irish and HE, and that the informational saliency of constituents is represented syntactically in both Irish and HE. The above examinations on the aspectual expressions of HE suggested the robust involvement of the sentence construction pattern in Irish. The verb-initial property of Irish has contributed to the formation of HE, inducing features apart from StE. The examination in this article specifically suggests that the *bí* construction in Irish has influenced the formation of aspectual expressions in HE, and that the *Is* construction has triggered the formation of expressions for information structure.

Based on the discussions so far, the following working hypothesis 1 is suggested: The main contributor of vocabulary (the lexifier) is English, and the basic morphosyntax also comes from English. Irish provides grammatical oppositions, thus serving as a base for expression formation.

In Section Irishness realized in English, we have examined the development of original forms in the system of HE. The distribution of *be after* was established in competition with other aspectual forms, one of which is the *have* perfect. In *do be V-ing*, habituality is marked as opposed to the *be* stative and the *be V-ing* continuous; furthermore, the Keane corpus has confirmed two types of *do be*, including *do be$_2$* in a relative clause. HE, as a matter of course, not only shows distinctive characteristics *vis-à-vis* StE, such as the three characteristics that the present study has highlighted, but also as a language, has an autonomous

grammar formation system in force, which produces expressions of tense/aspect and information structure.

With hypothesis 1, we may find some grammatical states of the supposedly stable HE (remember stage II in Figure 1) explicable, whereas the latest states of HE, which are relentlessly exposed to other varieties of English, remains unexplained. We must also take this exposure to other varieties and its effects into account. In fact, we have already come across some aspects in which the exposure seems to have influenced HE in prior sections of the article, including the extra use of *that* to mark informational saliency (The *'tis…* construction). Besides, Shimada (2013, 2016a) reports social connotations of *do be* among speakers of SwHE and illustrates speakers' awareness toward "Standard" based on her field research. Speakers of HE, in exposure to standard norms, find the form of *do be* bad grammar, sometimes with the stigma of "poorness" and "uneducatedness". The linguistic form has, in a way, served as a criterion of education and socioeconomic status.

Norm shift by the second contact

The *do be* form is an overt example of the second contact, where the first was between English and Irish to form HE (refer to Figure 1). The second contact is concerned with HE as a stable variety and other major varieties of English resulting from the spread of mass media, mainly television broadcasting, joining the EC, and increased access to education after the 1970s. In other words, in the second contact, HE went from endo-normative to exo-normative as interacting with main standard varieties of English. This norm shift is diagnosed by the acute rise of awareness of a "Standard" among speakers of SwHE.

Note that not all the Irish characteristics of HE are judged as bad grammar or taken negatively by the speakers. See details in (Shimada, 2010, 2015, 2016b), who has conducted a survey of speakers' awareness toward 26 HE sentences in 11 feature categories. The contrast in speaker perception of *do be* with *be after* is telling. While both *do be* V-*ing* and *be after* V-*ing* are equally regarded as showing Irishness, the *do be* form is judged as bad grammar and is avoided in speech while the *be after* form has a significantly more positive judgment. Shimada (2016a) explains the gap in their awareness with the term of morphosyntactic conformity. The sequence of two auxiliaries, *do* and *be*, does not comply with a morphosyntactic constraint of "Standard", which is an abstract construct in the linguistic knowledge of the speakers[18]. The *be after* form, on the other

---

18   As an anonymous reviewer has pointed out, *do* plus a non-finite verb does not generally violate morphosyntactic conformity when *do* is used emphatically. However, emphatic *do* does not cooccur with *be*, hence, i.e., **I DO be writing* is not registered by native speakers of English. As

hand, having two types of the complement V-*ing* and NP, does not violate English morphosyntax, although it is distinct from StE. The noun categories including gerund follow a preposition; this conforms precisely to English morphosyntax. Speakers of SwHE are highly conscious of morphosyntactic conformity but not semantic congruity. A form that satisfies morphosyntactic conformity is generally accepted.

The norm adopted in the primary and stable stages of contact-induced language formation, stages I and II in Figure 1, is not equivalent to the one that the lexifier language adopts for itself. HE was generated in the indigenously established norm by untutored adult L2 acquisition. The majority of Irish adult monoglots became bilingual through contact and interaction with other Irish people who had already acquired English in a context where English was a significant means to promote socioeconomic security. In such non-institutional contact situations, HE, whose speakers were in a process of shift from Irish to English, would be likely to exhibit transfer of Irish-driven sentence construction as well as culturally essential Irish lexicon. This state if agreed finds some similarities to the "endonormative stabilization" described by Schneider (2007) as one of the five characteristic stages in his Dynamic Model to capture general features of postcolonial English.[19] However, in a later stage, speakers of HE are assumed to have put into use a model of StE and have also acquired this through a written medium. The shift from a norm indigenously established on the community basis to an exo-normative model can be referred to as "norm shift".

Working hypothesis 2: In the process of HE formation to date, a norm shift occurred during the second contact (stage III in Figure 2). HE shifted its norm from the indigenously established one to a perceived Standard English. This has been triggered by contact with other major varieties of English since the 1970s (refer to Figure 1).

Based on speakers' statements from sessions with them during fieldwork (e.g., Shimada, 2013), it seems that the overall degree of speaker awareness of "Standard" grows when the variety they speak is undergoing contact with major or standardized varieties of the lexifier language. This awareness may affect the linguistic repertoires that speakers use in their everyday linguistic exchanges. The increased

---

such, the use of *do be* in HE clearly digresses from StE. Note that in habitual *do be* seen in HE, *do* is not accented, in contrast to emphatic *do*.
19 Schneider (2007), in his research on the evolution of postcolonial forms of English, proposes a five-stage process of emergence: foundation, exonormative stabilization, nativization, endonormative stabilization, and differentiation. Taking a general approach, Schneider (2007) does not focus on the linguistic process alone, instead describing general features of sociopolitical background, identity constructions, sociolinguistic conditions, and linguistic effects. Schneider (2007) considers HE a dialect of English, not a postcolonial English.

degree of awareness of "Standard" among speakers of HE can be an outcome of the norm shift. Contact with globally dominant or internationally standardized varieties is not special in the case of HE but is true of all languages that are currently exposed to a dominant variety of their lexifier language. Potential examples include African American Vernacular English (AAVE), Jamaican English, and TokPisin. Although of course careful research is needed before any definite claim can be made, AAVE and Jamaican English are exposed to mainstream American English and are clear cases of decreolization. In this sense, there seems to be similarities with HE in terms of the contact phenomenon. Norm shift may be another aspect that HE shares with other contact languages, including creoles exposed to the lexifier language in its matured stage.
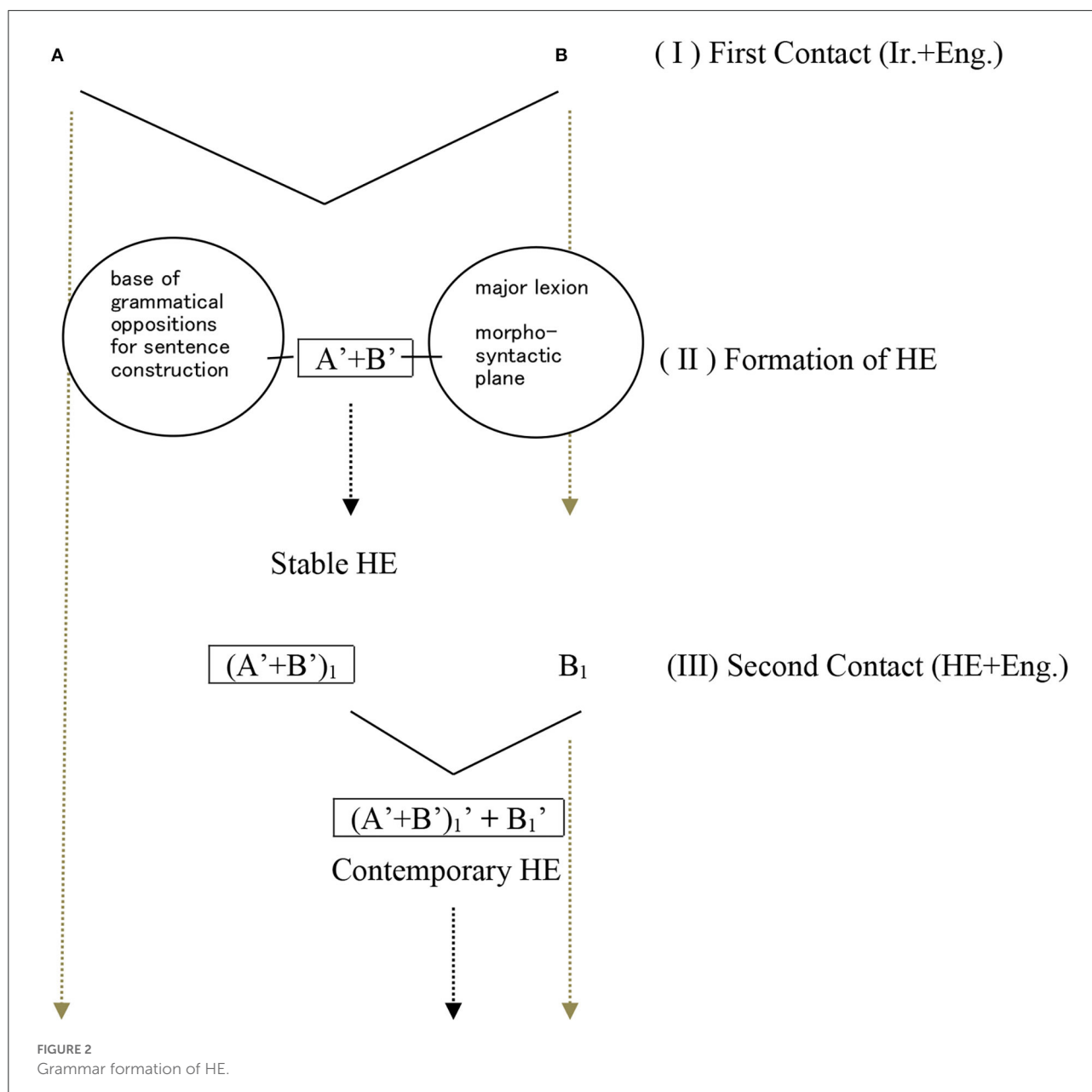
A model to be suggested

HE provides an example of what happens when two different languages come into contact within a community. HE, unlike creole languages, has a single substrate language. This has an advantage in isolating a specific area of substrate influence and assessing contact-induced grammatical formation. Harris (1984, p. 191) states: "the task of isolating specific areas of substratal influence would appear to be much easier in the case of contact vernaculars for which a single substratal language can be unambiguously identified. Varieties of this type would thus seem to constitute one of the most fruitful research sites for those interested in testing substratum hypotheses. One such vernacular is Irish English". HE thus provides a valuable material for investigating how the grammar of a new language is formed when two languages (we will call them "Language A" and "Language B" for convenience) meet, and for investigating contact-induced language change. Based on the examination of HE, this article concludes by proposing a model of the way grammar comes to be, stabilizes, and then undergoes further changes.

In Figure 2, A is given for Irish and B for English. In the first contact between A and B in a speech community (stage I), new language formation is facilitated, with the native speakers of A acquiring and accommodating B. The newly developed language is tagged with A'+ B' in the figure (stage II); this language forms a grammatical system and attains what can be described as "stable HE," which is thought to have been spoken by a few generations after the first contact between Irish and English. Stable HE is exemplified in varieties of southwest HE in the early to mid-20th century. The formation and development of HE is contingent upon the sociolinguistic circumstance that the history of Ireland has laid, sketched in Figure 1. As the working hypotheses 1 and 2 suggested in Where today's form came from and Norm shift by the second contact, in the case of HE, Irish offers a

FIGURE 2
Grammar formation of HE.

base for grammatical oppositions, while English provides the basic vocabulary and morphosyntax. The formation toward the stability of the system then proceeds.

Going further down the timeline, HE is then in the second contact phase and exposed to the major varieties of English. This is indicated by (III), where $(A' + B')_1$, the progressed version of $A' + B'$, faces intense contact with B_1 in the given community. The contact fosters convergence; the renewed language is tagged with $(A' + B')_{1'} + B_1$. This is the result of the second contact with another English variant B_1, experiencing a norm shift as suggested in the working hypothesis 2. Note

that following the process outlined in Figure 2, the element of A may be diluted (i.e., buried under new elements) but will never completely disappear even if further intensive contact with other varieties of English occurs, because it was part of the formation. To give an extreme example, in the distant future, we may end up with $((((A' + B') + B'') + C) + D)$, and some of the new elements may compete with A', producing complicated results that are exceedingly difficult to trace back to A', but A' is still there at the base, unless a drastic demographic change overtakes the community and a language replacement occurs. The heritage from Language A gradually

fades through a process of convergence with Language B, but even so, grammatical oppositions of Language A are retained in some way or another.

Conclusion

This article, taking salient characteristics of HE, including the *be-after* perfect, the *do be* habitual, and the *'tis…..* construction as windows to its underlying properties, has examined how the current forms and meanings of these characteristics are the result of a dynamic process of language formation guided by language contact. The article surveyed earlier forms in the rise of HE, described some facets of contemporary HE, and finally introduced an integrated perspective across the characteristics to discuss how elements from Irish have been inherited and realized in English and updated over time until the present.

One of the main issues was Irish language traits and their realization in HE morphosyntax. Examinations of expressions of tense/aspect and information structure in HE have led to the conclusion that it has inherited grammatical oppositions from Irish. The division of grammatical categories, with which states of affairs are rendered, has been realized using English morphosyntax. This article illustrated cases of habituality marking and saliency marking, in which HE adopts the oppositions from Irish. Also pointed out was the high degree of involvement of the Irish sentence construction pattern in the aspectual expressions of HE, wherein the verb-initial property of Irish has contributed to the formation of HE. Specifically, this article argues that the *bí* construction of Irish has influenced the formation of aspectual expressions in HE, and that the *Is* construction of Irish has triggered the formation of expressions for information structure. These findings suggest that further investigation of the Irish morphosyntax is likely to show even more involvement of Irish in the formation of HE.

From a sociolinguistic perspective, norm shift, which occurs in the second contact with other varieties of English (refer to Figure 2), presumably over the past 50 years or so, is key in the explanation why HE is as it is today. Speakers' awareness toward the standard norm brings them to prune HE so that it conforms to the morphosyntactic constraints of StE. Importantly, the model presumes that grammatical oppositions of Irish remain even after such pruning, although the heritage from Irish is gradually diluted through a process of convergence with other varieties of English. This presumption and the supposed model should be tested against many different grammatical forms of phenomena in HE in the future both to assume this model's variability and to uncover the nature of the forms and phenomena.

This article has confined itself to the description of HE; thus, the model suggested within addresses the grammar formation of HE, considering settings and language contact situations particular to HE. However, the model, along with the two working hypotheses of the formation of grammar in a contact situation and of norm shift, can be applied to find similarities and contrasts with other language-contact phenomena. This, in turn, can contribute to various studies in a variety of languages, helping to enrich our understanding of contact-induced language change and the mechanism of grammar formation of a language in contact.

Data availability statement

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author.

Author contributions

TS: design of the study, collection and analysis of the data, and preparation of the manuscript.

Funding

Acknowledgments

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amador-Moreno, C., McCafferty, K., and Vaughan, E. (2015). *Pragmatic Markers in Irish English*. Amsterdam: John Benjamins. doi: 10.1075/pbns.258

Barron, A., and Schneider, K. P. (2005). *The Pragmatics of Irish English*. Berlin: Mouton de Gruyter. doi: 10.1515/9783110898934

Bliss, A. J. (1972). Languages in contact: some problems of Hiberno-English. *Proc. R. Ir. Acad. Sci.* 72, 63–82.

Bliss, A. J. (1979). *Spoken English in Ireland 1600-1740*. Dublin: The Dolmen Press.

Comrie, B. (1976). *Aspect*. Cambridge: Cambridge University Press.

Cotter, C. (1994). "Focus in Irish and English: contrast and contact," in *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistic Society*. p. 134–144. doi: 10.3765/bls.v20i1.1450

Curme, G. O. (1931). *A Grammar of the English Language, Volume II syntax*. Essex: Verbatim.

DeGraff, M. (2005). Morphology and word order in "creolization" and beyond. In *The Oxford Handbook of Comparative Syntax*, Cinque, G., and Kayne, R. S., (eds). Oxford: Oxford University Press. p. 293–372.

Dillon, M. (1941). Modern Irish atá sé déanta agam 'I Have Done It'. *Language*. 17, 49–50. doi: 10.2307/409460

Doyle, A. (2001). *Irish*. Muenchen: Lincom Europa.

Edwards, R. D. (1981). *An Atlas of Irish History, Second edition*. London: Routledge.

Ellgård, A. (1953). *The Auxiliary do: The Establishment and Regulation of its Use in English*. Stockholm: Almqvist and Wiksell.

Filppula, M. (1986). *Some Aspects of Hiberno-English in a Functional Sentence Perspective, University of Joensuu Publications in the Humanities 7*. Joensuu: University of Joensuu.

Filppula, M. (1990). Substratum, superstratum, and universals in the genesis of Hiberno- English. *Irish Uni. Rev.* 20, 41–54.

Filppula, M. (1999). *The Grammar of Irish English: Language in Hibernian Style*. London: Routledge. doi: 10.2307/29742738

Filppula, M., Klemola, J., and Paulasto, H. (2008). *English and Celtic in Contact*. London: Routledge. doi: 10.4324/9780203895009

Finegan, E. (1995). "Subjectivity and subjectivisation: An introduction," in: *Subjectivity and Subjectivisation,* Stein, D., and Wright, S. (eds). Cambridge: Cambridge University Press 1–10. doi: 10.1017/CBO9780511554469.001

FitzGerald, G. (1984). Estimates for baronies of minimal level of Irish speaking amongst successive decennial cohorts: 1771-1781 to 1861-1871. *Proc. R. Ir. Acad. Sci.* 84, 117–155.

Greene, D. (1966). *The Irish Language*. Dublin: The Cultural Relations Committee of Ireland.

Greene, D. (1979). Perfect and perfectives in Modern Irish. *ÉRIU*. 30, 122–141.

Guilfoyle, E. (1985). Hiberno-English: A parametric approach. In Perspectives on the English Language in Ireland: Proceedings of the First Symposium on Hiberno-English, Harris, J., Little, D., and Singleton, D. Dublin: Centre for Language and Communication Studies, Trinity College Dublin. P. 121–132.

Harris, J. (1984). Syntactic variation and dialect divergence. *J. Linguistics.* 20, 303–327. doi: 10.1017/S0022226700013918

Harris, J. (1986). Expanding the superstrate: Habitual aspect markers in Atlantic Englishes. *English World-Wide.* 7, 171–199. doi: 10.1075/eww.7.2.02har

Harris, J. (1990). More on brogues and creoles: What's been happening to English short *u*? *Irish Univ. Rev.* 20, 73–90.

Harris, J. (1991). "Conservatism versus Substratal Transfer in Irish English," in *Dialects of English: Studies in Grammatical Variation*, Trudgill, P., and

Chambers, J. K. (edds). London: Longman. p. 191–212. doi: 10.4324/9781315505459-14

Harris, J. (1993). "The grammar of Irish English," in *Real English: The Grammar of English Dialects in British Isles*, Milroy, J., and Milroy, L. London: Longman. p. 139–186.

Hartmann, H. (1974). Distribution und Funktion der Expanded Form in einigen Dialekten der irischen Sprache von Co. Galway. *Zeitschrift für celtische Philologie.* 33, 140–284. doi: 10.1515/zcph.1974.33.1.140

Heine, B., and Kuteva, T. (2005). *Language Contact and Grammatical Change*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511614132

Henry, A. (1995). *Belfast English and Standard English: Dialect Variation and Parameter Setting*. Oxford: Oxford University Press.

Henry, P. L. (1957). *An Anglo-Irish Dialect of North Roscommon*. Dublin: University College Dublin.

Hickey, R. (2000). "Models for describing aspect in Irish English," in *The Celtic Englishes* II, Tristram, H. L. C. (ed). Heidelberg: Carl Winter. p. 97–116.

Hickey, R. (2007). *Irish English: History and Present-day Forms*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511551048

Hickey, R. (2016). *Sociolinguistics in Ireland*. Basingstoke: Palgrave Macmillan. doi: 10.1057/9781137453471

Hickey, R., and Amador-Moreno, C. P. (2020). *Irish Identities: Sociolinguistic Perspectives*. Berlin, New York: De Gruyter. doi: 10.1515/9781501507687

Joyce, P. W. (1910). *English as We Speak It in Ireland*. Dublin: Wolfhound Press.

Kallen, J. L. (1985). Bilingualism and the genesis of Hiberno-English syntax. *TEANGA.* 5, 98–113.

Kallen, J. L. (1986). "The co-occurrence of do and be in Hiberno-English syntax," in *Perspectives on the English Language in Ireland: Proceedings of the First Symposium on Hiberno-English*, Harris, J., Little, D., and Singleton, D. Dublin: Centre for Language and Communication Studies Trinity College Dublin. p. 133–147.

Kallen, J. L. (1989). Tense and aspect categories in Irish English. *English World-Wide.* 10, 1–39. doi: 10.1075/eww.10.1.02kal

Kallen, J. L. (1990). The Hiberno-English perfect: grammaticalisation revisited. *Irish University Rev.* 20, 120–136.

Kallen, J. L. (1991). "Sociolinguistic variation and methodology: *after* as a Dublin variable," in *English Around the World: Sociolinguistic Perspectives*, Cheshire, J. Cambridge: Cambridge University Press 61–74. doi: 10.1017/CBO9780511611889.005

Kallen, J. L. (1994). "English in Ireland," in *English in Britain and Overseas: Origins and Development, vol. V, The Cambridge History of the English Language,* Robert W. (ed). Burchfield, Cambridge: Cambridge University Press. p. 148–196. doi: 10.1017/CHOL9780521264785.005

Kallen, J. L. (1997). *Focus on Ireland*. Amsterdam: John Benjamins. doi: 10.1075/veaw.g21

Kallen, J. L. (2013). *Irish English Volume 2: The Republic of Ireland*. Berlin: De Gruyter. doi: 10.1515/9781614511298

Kallen, J. L., and Kirk, J. M. (2007). "Accessing Celticity in a corpus of Irish standard English," in *The Celtic Languages in Contact*, Tristram, H. L. C. (ed). Potsdam: Potsdam University Press 270–288.

Kallen, J. L., and Kirk, J. M. (2008). *The International Corpus of English*. Ireland: CD-ROM.

Keane, J. B. (1959). *Sive*. Dublin: Progress House.

Keane, J. B. (1961). *Highest House on the Mountain*. Cork: Mercier Press.

Keane, J. B. (1967). "Letters of a Successful TD," in *The Celebrated Letters of John B. Keane*. Cork: Mercier Press.

Kirk, J. M. (2017). The present perfect in Irish English. *World Englishes.* 36, 239–253. doi: 10.1111/weng.12256

Mac Eoin, G. (1993). "Irish," in *The Celtic Languages*, Martin, J. (ed). Ball with James Fife, London: Routledge. p. 101–144.

McCafferty, K. (2017). Irish English in emigrant letters. *World Englishes.* 36, 176–190. doi: 10.1111/weng.12252

McCafferty, K., and Amador-Moreno, C. P. (2012). "CORIECOR: A corpus of Irish English correspondence: Compiling and using a diachronic corpus to study the evolution of Irish English," in *New Perspectives on Irish English*, Migge, B., and Ní Chiosá, N., Amsterdam: John Benjamins. p. 265–287. doi: 10.1075/veaw.g44.13mcc

McCawley, J. (1971). "Tense and time reference in English," in *Grammar and Meaning*, McCawley, J. London: Academic Press. p. 257–272.

McCloskey, J. (1979). *Transformational Syntax and Model Theoretic Semantics: A Case Study in Modern Irish*. Dordrecht: D. Reidel. doi: 10.1007/978-94-009-9495-9

McCloskey, J. (1996). "Subjects and subject positions in Irish," in *The Syntax of the Celtic Languages: A Comparative Perspective,* Borsley, R. D., and Roberts, I. (eds). Cambridge: Cambridge University Press. 241–283. doi: 10.1017/CBO9780511586279.009

McCloskey, J. (2005). "A note on predicates and heads in Irish clausal syntax," in *Verb first: On the Syntax of Verb-initial Languages,* Carnie, A., Harley, H., and Dooley, S. A. Amsterdam: John Benjamins. p. 155–174. doi: 10.1075/la.73.12mcc

Murphy, B. (2006). Gender and its influence on linguistic production in responding across different age groups. Presented at *Sociolinguistic Symposium* 16, University of Limerick.

Nolan, B. (2006). The passives of modern Irish. In *Passivization and Typology: Form and Function*, eds. A. Werner and L. *Leisiö, Amsterdam: John Benjamins* 132–164. doi: 10.1075/tsl.68.10nol

Ó Cuív, B. (1986). "Irish language and literature, 1691-1845," in *A New History of Ireland: IV Eighteenth-century Ireland 1691-1800*, Moody, T. W., and Vaughn, W. E. (eds). London: Clarendon Press. p. 374–423.

Ó Dochartaigh, C. (1992). "The Irish language," in *The Celtic Languages*, MacAulay,D. (ed). Cambridge: Cambridge University Press. p. 11–99. doi: 10.1017/CBO9780511551871.003

Ó hÚrdail, R. (1997). "Hiberno-English: Historical background and synchronic features and variation", in *The Celtic Englishes*, Tristram, H. L. C. (ed). Heidelberg: Universitätsverlag C. *Winter* 180–200.

Ó Sé, D. (2004). The 'after' perfect and related constructions in Gaelic dialects. *ÉRIU.* 54, 179-251. doi: 10.1353/eri.2004.0015

Ó Siadhail, M. (1989). *Modern Irish: Grammatical Structure and Dialectal Variation.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511620621

Odlin, T. (1997). Hiberno-English: pidgin, creole, or neither? *CLCS Occasional Paper.* 49, 1-39. Dublin: Trinity College Dublin.

Rissanen, M. (1991). "Spoken language and the history of do-periphrasis," in *Historical English Syntax,* Kastovsky, D. (ed). Berlin: Mouton de Gruyter.

Ronan, P. (2005). "The *after*-perfect in Irish English," in *Dialects Across Borders*, Filppula, M., Klemola, J., Palander, M., and Penttilä, E, Amsterdam: John Benjamins. p. 253–270. doi: 10.1075/cilt.273.16ron

Russell, P. (1995). *An Introduction to the Celtic Languages*. London: Longman.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the World.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511618901

Schrier, A. (1997). *Ireland and the American Emigration 1850–1900, Second edition.* Chester Springs PA: Dufour Editions.

Shimada, T. (2008). Airurando eigo *be after V-ing* no hyougen kouka: *have* kanryou to no tairitsu wo chuushin ni [On the Pragmatics of the *be after* Construction in Hiberno-English: a Contrast with the *have* Perfect]. *Tokyo University Linguistic Papers.* 27, 187–206.

Shimada, T. (2010). *English in Ireland: Beyond Similarities*. Hiroshima: Keisuisha.

Shimada, T. (2013). The *do be* form in southwest Hiberno-English and its linguistic enquiries. *Festschrift for Professor Hiroshi Kumamoto, Tokyo University Linguistic Papers.* 33, 255–271. doi: 10.15083/00027516

Shimada, T. (2015). Morphosyntactic features in flux: awareness of 'Irishnessness' and 'Standard' in Hiberno-English speakers. *Arbbeiten aus Anglistik und Amerikanistik.* 40, 41–65. Available online at: https://www.jstor.org/stable/24922039

Shimada, T. (2016a). Speakers' awareness and the use of *do be* vs. *be after* in Hiberno-English. *World Englishes.* 35, 310–323. doi: 10.1111/weng.12197

Shimada, T. (2016b). *Eigo to iu Sentaku: Airurando no Ima [lit. The Choice of English: Ireland Today]*. Tokyo: Iwanami.

Shimada, T. (2018). '*Tis....* pattern in Hiberno-English as a grammatical innovation. *Festschrift for Professor Tooru Hayasi, Tokyo University Linguistic Papers.* 39, 243–263. doi: 10.15083/00074543

Siemund, P. (2004). "Substrate, superstrate and universal. Perfect constructions in Irish English," in *Dialectology Meets Typology,* Kortmann, B. (ed). Berlin: Mouton de Gruyter. 401–434.

Smith, G., and Hickey, D. (2002). *John B.* Cork: The Mercier Press.

Stenson, N. (1981). *Studies in Irish Syntax*. Tübingen: Gunter Narr.

Stenson, N. (2020). *Modern Irish: A Comprehensive Grammar*. London: Routledge. doi: 10.4324/9781315302034

Taniguchi, J. (1972). *A Grammatical Analysis of Artistic Representation of Irish English: With a Brief Discussion of Sounds and Spelling*. Tokyo: Shinozaki Shorin.

Thurneysen, R. (1946). *A Grammar of Old Irish, Translated From the German by D. A. Binchy and O. Bergin*. Dublin: School of Celtic Studies, Dublin Institute for Advanced Studies.

Todd, L. (1999). *Green English*. Dublin: The O'Brien Press.

Tristram, H. L. C. (1997). *The Celtic Englishes*. Heidelberg: Carl Winter Universitätsverlag.

van Hamel, A. G. (1912). On Anglo-Irish syntax. *Englische Studien.* 45, 272–292.

Winford, D. (2001). " "Intermediate" creoles and degrees of change in creole formation: The case of Bajan," in *Degrees of Restructuring in Creole Languages,* Neumann-Holzschuh, I., and Schneider, E. W. Amsterdam: John Benjamins. 215–246. doi: 10.1075/cll.22.13win

Winford, D. (2009). "The interplay of 'Universals' and contact-induced change in the emergence of New Englishes," in *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond,* Filppula, M., Klemola, J., and Paulasto, H. London: Routledge. p. 206–230.

Wright, J. (1900). *English Dialect Dictionary*. Oxford: Oxford University Press.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership