



# The Building Blocks of Child Bilingual Code-Mixing: A Cross-Corpus Traceback Approach

Antje Endesfelder Quick<sup>1\*</sup> and Stefan Hartmann<sup>2</sup>

<sup>1</sup> Faculty of Philology, Institute of British Studies, University of Leipzig, Leipzig, Germany, <sup>2</sup> Faculty of Arts and Humanities, German Department, University of Düsseldorf, Düsseldorf, Germany

This paper offers an inductive, exploratory study on the role of input and individual differences in the early code-mixing of bilingual children. Drawing on data from two German-English bilingual children, aged 2–4, we use the traceback method to check whether their code-mixed utterances can be accounted for with the help of constructional patterns that can be found in their monolingual data and/or in their caregivers' input. In addition, we apply the traceback method to check whether the patterns used by one child can also be found in the input of the other child. Results show that patterns found in the code-mixed utterances could be traced back to the input the children receive, suggesting that children extract lexical knowledge from their environment. Additionally, tracing back patterns within each child was more successful than tracing back to the other child's corpus, indicating that each child has their own set of patterns which depends very much on their individual input. As such, these findings can shed new light on the interplay of the two developing grammars in bilingual children and their individual differences.

**Keywords:** corpus linguistics, traceback method, usage-based linguistics, code-mixing, individual differences

## OPEN ACCESS

### Edited by:

Marianne Gullberg,  
Lund University, Sweden

### Reviewed by:

Virve-Anneli Vihman,  
University of Tartu, Estonia  
Marilyn Vihman,  
University of York, United Kingdom

### \*Correspondence:

Antje Endesfelder Quick  
antje.quick@uni-leipzig.de

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Psychology

**Received:** 19 March 2021

**Accepted:** 14 May 2021

**Published:** 27 July 2021

### Citation:

Quick AE and Hartmann S (2021) The  
Building Blocks of Child Bilingual  
Code-Mixing: A Cross-Corpus  
Traceback Approach.  
Front. Psychol. 12:682838.  
doi: 10.3389/fpsyg.2021.682838

## INTRODUCTION

Usage-based accounts of linguistic phenomena have become increasingly important in the field of language acquisition. In this paper, we apply this account to bilingual language acquisition, and more specifically to utterance-internal child bilingual code-mixing, i.e., the simultaneous use of two languages within one utterance, e.g., *aber bloß a little bit* “but just a little bit” (Silvie, 3;07)<sup>1</sup>. Code-mixing, or code-switching<sup>2</sup>, has been a recurrent topic both in sociolinguistics and in psycholinguistics (see Bullock and Toribio, 2009; Gardner-Chloros, 2009 for an overview). Most research on code-mixing so far (e.g., Poplack, 1980; Myers-Scotton, 1997; MacSwan, 2000; Cantone, 2007) has adopted a structuralist/constraint-based framework. Both formalist and usage-based approaches are of course more heterogeneous in terms of theory and methodology than this simplified, coarse-grained division might suggest, and there is also a considerable degree of overlap between them (see e.g., Yang, 2016). As a general tendency, however, it seems fair to say that most formalist proposals focus more on structural constraints on code-mixing, while usage-based accounts are more interested in the cognitive mechanisms that underlie code-mixing. From a

<sup>1</sup> Throughout the paper, German material is presented in boldface. All examples from the corpora described below.

<sup>2</sup> While these terms are often used interchangeably, we follow Muysken (2000) in using the term “code-mixing” for utterance-internal code-switching.

usage-based perspective, it can be assumed that the same mechanisms that have been shown to drive monolingual acquisition play a role in multilingual acquisition and code-mixing as well. These mechanisms include analogy and pattern finding (Tomasello, 2009) as well as chunking and frequency effects (see e.g., Diessel, 2019).

Recent studies have therefore started to investigate code-mixing from a usage-based perspective (e.g., Quick et al., 2018; Vihman, 2018). In this study, we follow up on this trend, focusing on the role of multi-word patterns as well as on individual differences between speakers. According to the usage-based approach to language acquisition, children acquire their linguistic knowledge based on their experience with the world and what they hear (e.g., Tomasello, 2003). Since no two children live the same life and each has different input situations, it stands to reason that variation in the output is the norm and each child has their own inventory of constructions. Monolingual children already exhibit an enormous range of variation. In multilingual speakers, variation is likely to be greater by virtue of being exposed to two languages and their respective interlocutors who speak different languages in different contexts. Multilingual speakers will frequently produce code-mixed utterances such as *der moon kann fly* “the moon can fly” (Fion, 3;2.12). In this paper, we discuss how bilingual code-mixing can be assessed in an exploratory, data-driven way, taking individual differences into account. To do so, we compare the code-mixing of two German-English bilingual children. First, we give a brief overview of the usage-based perspective on (monolingual as well as multilingual) language acquisition and code-mixing before we turn to our corpus-based case study, in which we inductively analyze the language use of two bilingual children on the basis of longitudinal corpus data.

## THEORETICAL PRELIMINARIES AND MAIN HYPOTHESES

The usage-based approach to language acquisition assumes that children acquire language by finding patterns in the input they receive (for an overview see Ambridge and Lieven, 2011; Ibbotson, 2020). Given the assumption that children structure their first words and utterances around their immediate experiences, it follows that their inventory of linguistic knowledge does not solely consist of words and grammar. Rather, the inventory is a mixture of lexically specific units (single words like *cat*, *dog*, as well as multiword expressions like *what’s this?*) and frame-and-slot patterns like [*what’s X?*] (see e.g., Tomasello, 2003: pp. 105–108). Multi-word units (MWUs), i.e., sequences of frequently co-occurring words, play a particularly important role in language acquisition. They can be acquired in different ways: Either the MWU is always encountered as such by the child and therefore not segmented and acquired as a whole, or the MWU emerges gradually through the frequent co-occurrence of certain words. But no matter how a unit was formed, it does not always have to stay a unit: Over time, children tend to break up multi-word units, e.g., using them as the basis for frame-and-slot patterns by opening a variable slot, and thus arriving at a

more productive use of their language (see e.g., Ambridge and Lieven, 2011: pp. 133–136). Consequently, the composition of inventories changes constantly and any description will always be a snapshot. Nevertheless, these snapshots are important because they tell us something about the ways children process and acquire language, as well as about the interplay of language and cognition.

Bilingual children are of particular interest because they can show patterns in their speech that are different from monolingual speech, such as code-mixing or other transfer phenomena (Koch and Günther, 2021). As mentioned in section Introduction, a large body of research has been accumulated (e.g., MacSwan, 2000; Bernardini and Schlyter, 2004; Cantone, 2007). While previous studies have acknowledged the existence of individual differences and distinguished different types of code-mixing, they mainly concentrated on the categorization of the various types of mixing and on describing constraints on code-mixing, linking them to potential underlying principles (e.g., Di Sciullo et al., 1986; Myers-Scotton and Jake, 2015). However, various studies have shown that these constraints are tendencies at best which very often cannot accommodate counterexamples, and have called for more flexible and dynamic models (e.g., Vihman, 2018; Backus, 2021).

Recently, a set of studies has taken code-mixing onto usage-based grounds suggesting that fixed chunks and frame-and-slot patterns that have been shown to play a major role in monolingual acquisition can also account for children’s code-mixing. Lexically fixed patterns make execution faster and less effortful because they are uttered without close monitoring (e.g., Havron and Arnon, 2021). On this view, code-mixing is suggested to be constructed around frame-and-slot patterns with the frame activated in one language and the open slot being filled with elements from the other language, e.g., [*that’s my* \_] + *Bademantel* “bathing gown” → *that’s my Bademantel* (Fion, 03;11.16). Quick et al. (2019) have shown that many of the patterns attested in one child’s code-mixing could be traced back to the caregivers’ input, suggesting that the child extracted patterns from the input to use it in his code-mixing. If we now compare the language of two children and their respective input, we should find individual differences in their use of patterns and their inventories. These differences can be expected to project into their code-mixing: Children make use of different patterns in their code-mixing. Code-mixing in bilingual children offers us a window into the complexities and interplay of the two developing grammars, which is why we will focus on code-mixed utterances in our corpus study to account for the individual inventories of the two children under scrutiny.

In order to do so, we need a reliable method to identify these inventories. In previous work, we have shown that the traceback method established by Lieven et al. (2003, 2009) and Dabrowska and Lieven (2005) for analyzing monolingual data is well-suited for identifying recurrent patterns in multilingual data as well (Quick et al., 2019, 2021). The basic idea of this method is to trace back all utterances in a test corpus to previous utterances based on a limited set of operations (see below for details). The traceback method was initially developed to substantiate the hypothesis that early child language is highly

formulaic and organized around a very limited set of “pivot schemas.” Indeed, the proportion of successful tracebacks proved to be very high consistently across all traceback studies. This in turn lends support to the position that children learn language from the input they receive, without any need for an innate “language acquisition device.” These results could be obtained across different languages, including German (Koch, 2019) and Italian (Miorelli, 2017), although it should be stressed that the way in which the traceback method operationalizes the detection of constructions works best in languages with a relatively fixed word order (see Miorelli, 2017; Koch, 2019; also see section Conclusion below). The method has also been used as a starting point for a more in-depth analysis of the constructional patterns that it retrieves. When studying code-mixing, the method can give clues to what extent children draw on frame-and-slot patterns that can also be found in their monolingual data, as well as in the input they receive. Thus, the use of the traceback method serves multiple complementary goals: Firstly, it allows us to quantify the extent to which a child’s early language use can be accounted for with a relatively simple set of fixed chunks and frame-and-slot patterns. Secondly, it allows us to identify those patterns, which also allows us to characterize each child’s individual inventory of constructions. This in turn can give clues as to the individual differences between children. Thirdly, the traceback method allows us to check to what extent the patterns in a child’s output overlap with patterns attested in the input they receive.

In this paper, we extend a previous study by Quick et al. (2019) by discussing what the traceback method can contribute to inductively identifying individual differences in children’s code-mixing. Our aim is to check (a) whether the code-mixed utterances can be constructed from the monolingual ones and (b) how much each child’s output correlates with the caregivers’ input. In addition, (c) we cross-correlate each child’s output with the input and the monolingual language use of the other child. Our main hypothesis is that there is a high degree of overlap between the patterns identified in the individual children’s language use, including their code-mixing, and those identified in their caregivers’ data. In addition, we expect that the proportion of successful tracebacks will be smaller for the cross-corpus than for the within-corpus studies due to the individual differences between the children and their input situation.

## CORPUS STUDY

### Participants

For the present study, we investigate two German-English bilingual children, Silvie and Fion. Both children grew up in one-parent-one-language (OPOL) households with one parent being a native speaker of English and the other being a native speaker of German. Both children lived in Germany, came from a middle-class household and are simultaneous German-English bilinguals. The two children were not acquainted with each other.

The first child, Silvie, had an English-speaking mother and a German-speaking father. The father’s proficiency in English was very limited and the parents therefore spoke German to one another. The corpus covers recordings from 2;4 until 3;10

years of age, averaging to about 2.5 h of recordings per week. For our analyses we included a total of 37,995 child utterances and 193,993 caregiver utterances.

Fion is the second child to a German-speaking mother and an English-speaking father. Although the parents mostly adhered to the OPOL strategy when they talked to Fion, they did not settle on a family language and sometimes used both languages interchangeably when all family members were present. Fion’s data covered a span from 2;3 to 3;11 and 47,812 child utterances as well as 180,293 input utterances entered the analyses. The input utterances include a small amount of data from Fion’s older brother, who also grew up as a simultaneous bilingual and sometimes used code-mixing when talking to Fion or his parents. The data were transcribed and enriched with a small set of annotations by the first author. For non-standard word forms, normalized forms were annotated (e.g., *gasbet* > *basket*, *de muffin* > *the muffin*, etc.). These normalized forms were also used for the traceback analysis.

### Corpus Analysis

To analyze the corpus data, we draw on the traceback method, which allows us to identify recurrent patterns in the data and which will be described in detail in the following section. In discussing the results, we additionally draw on an exploratory analysis of word pairs (bigrams) attested in the code-mixed data. Taken together, these approaches can help us detect the “building blocks” of code-mixed utterances in the early speech of bilinguals.

### The Traceback Method

We follow Quick et al. (2019), who used a variant of the traceback method established in seminal works like Lieven et al. (2003) and Dabrowska and Lieven (2005). In the traditional application of the method, a longitudinal corpus documenting the language acquisition of one child is split into two parts: the test corpus, which usually contains the last two recording sessions, and the main corpus, which contains all previous recordings. The goal of the method is to show that the vast majority of the child’s utterances in the test corpus have predecessors in the main corpus, i.e., they are either verbatim repetitions (called “fixed strings”), or they can be accounted for with the help of “templates” that are partially lexically specific and contain an open slot, such as [*I want* REFERENT]. In addition, the method can help us answer the question which patterns the child uses.

Quick et al. (2019) deviate from the traditional application in the way they carve up their dataset into main and test corpus: Investigating the code-mixing of Fion, they use the child’s code-mixed data as test corpus and the child’s own as well as his caregivers’ monolingual utterances as main corpus. In this way, they show that even most of the child’s code-mixing can be accounted for on the basis of partially filled constructions. This suggests that in essence, the same patterns that account for children’s monolingual language use can also account for their code-mixing. In the present study, we extend this analysis, combining it with a cross-corpus traceback approach as proposed in Koch et al. (2020). While we only relied on utterance-initial *n*-grams in Quick et al. (2019), the computational implementation of the traceback algorithm in this paper is closer to the original

traceback method, although it is still simplified in order to allow for a fully automatic analysis<sup>3</sup>. For the present study, we used an algorithm that works as follows (see data availability statement for more detailed information):

1. For each utterance in the test corpus, it checks whether there is a verbatim match in the main corpus. If there is a match, the derivation is considered successful.
2. If there is no match in the main corpus, it checks whether a frame-and-slot pattern can account for the utterance. To do so, up to two consecutive words are replaced by a wildcard in the search expression (equivalent to the SUBSTITUTE operation in the classic traceback procedure). For example, if our target utterance is *das hat time out* “this has time out” (Fion, 02;03.12), the algorithm will check if attestations of *\_\_ hat time out*, *das \_\_ time out*, *das hat \_\_ out*, *das hat time \_\_*, *das hat \_\_ \_\_ time out*, *das \_\_ out* are attested in the corpus at least twice (the threshold established by Dabrowska and Lieven, 2005). Then the algorithm checks if the omitted words (e.g., *das* in the case of *\_\_ hat time out*) are attested in the main corpus. Only if this is the case, the pattern candidate is treated as valid. If there are multiple valid pattern candidates, the ones with the longest consecutive fixed string are preferred, e.g., *das hat \_\_* (two consecutive words in the fixed string) is preferred over *das \_\_ out* (only one word before and after the open slot). Also, pattern candidates with utterance-initial fixed strings are preferred over candidates with an utterance-initial open slot: Given the tendency toward right-headedness in both English and German and given the results of earlier studies (see e.g., Cameron-Faulkner et al., 2003 on the relevance of utterance-initial elements in child-directed speech), this promises more plausible results. However, the rule of longest consecutive strings is prioritized over the rule to prefer utterance-initial patterns. If no pattern candidate fulfills the requirements (at least two occurrences in the main corpus, and the omitted words have to be attested in the main corpus as well), then the derivation is considered unsuccessful.

We used the code-mixed utterances ( $N = 3,501$  for Fion and 4,279 for Silvie) as test corpus and (a) the child’s own monolingual utterances and (b) the caregivers’ data as main corpora. In a second step, we used (c) the *other* child’s monolingual data and (d) the *other* child’s input as the main corpus. We refer to (a) and (b) as within-corpus traceback and to (c) and (d) as cross-corpus traceback. Using the within-corpus approach, we check to what extent the children’s code-mixed utterances can be accounted for with the help of fixed chunks and frame-and-slot patterns that they have either used themselves or that they have heard in their input. The cross-corpus approach can help us to get a clearer impression of the extent to which the linguistic repertoires of the two children differ. Compared to other implementations of the traceback method, our approach entails the disadvantage that the pattern detection process does not take semantic and/or syntactic information into account, which can lead to rather implausible patterns being postulated. However, there is no guarantee that the

linguistically informed patterns identified in previous traceback studies are psychologically plausible (see e.g., Hartmann et al., 2021). The purely data-driven approach can also be seen as an advantage as it detects patterns purely on the surface level without making far-reaching a-priori assumptions.

## Results

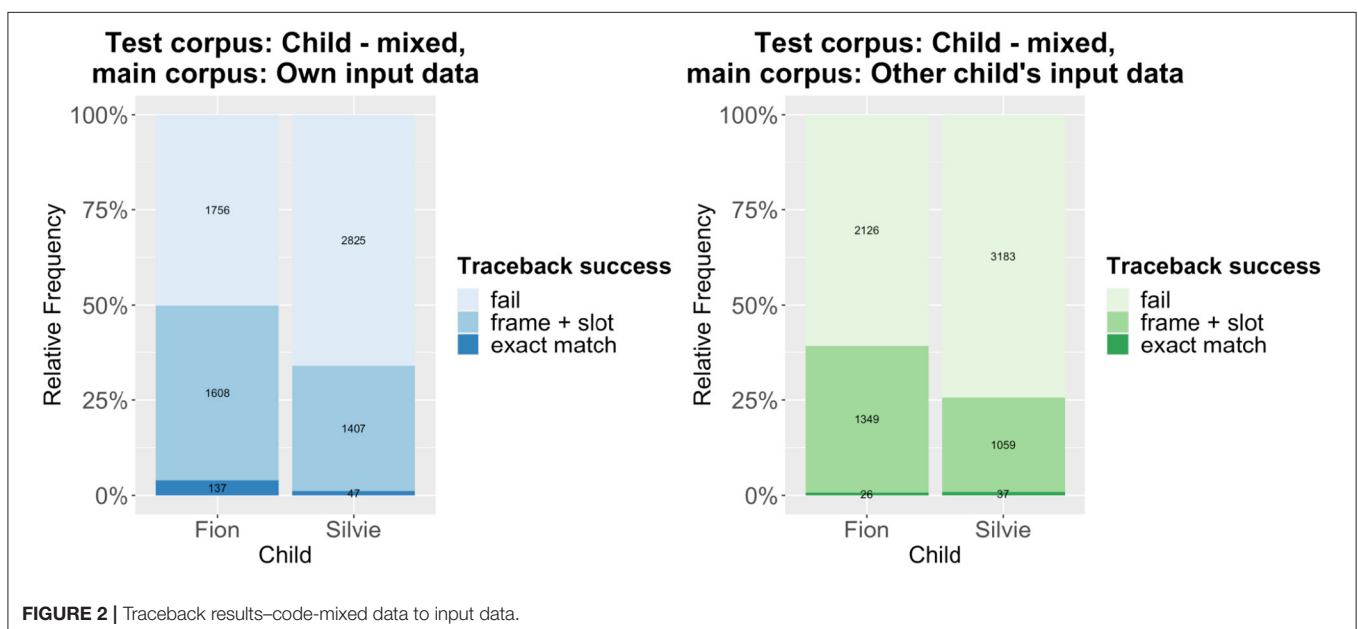
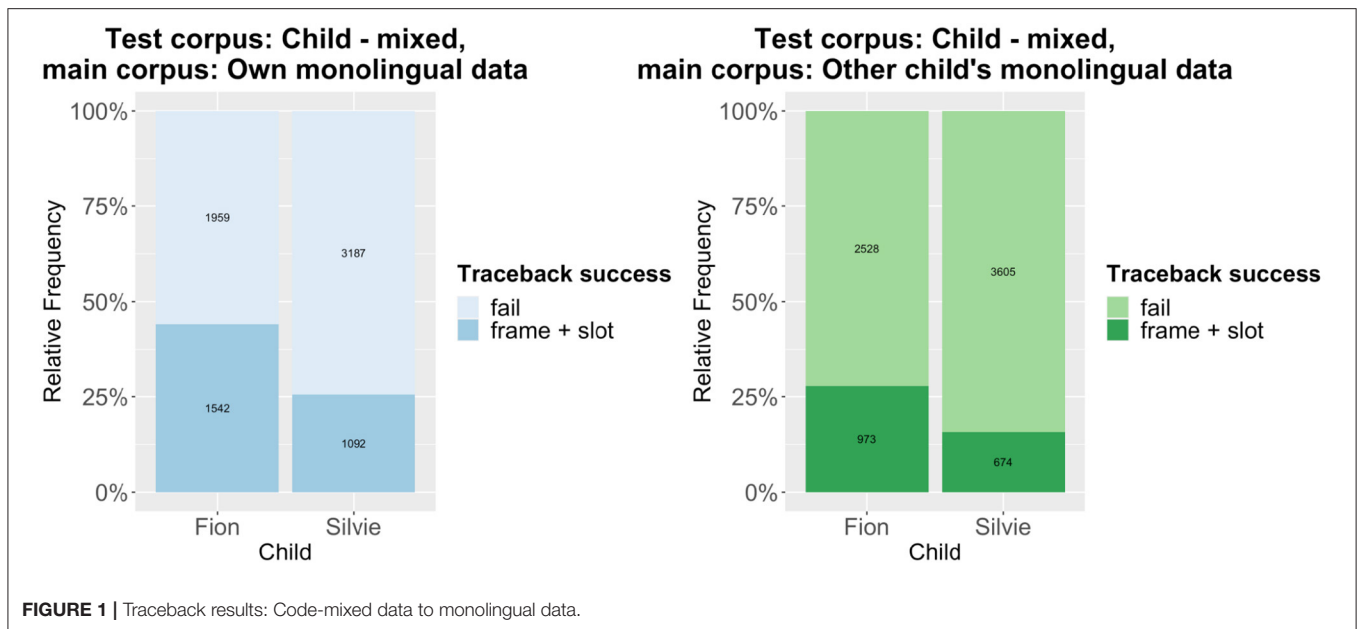
The traceback results are summarized in **Figures 1, 2**. **Figure 1** shows the results of tracing the code-mixed data to the monolingual data, while **Figure 2** shows the results that are obtained when using the caregivers’ input as main corpus. Compared to other traceback studies, the success rate is comparatively low. However, we have to remember that the test corpora only include code-mixed utterances, while the main corpora almost exclusively consist of monolingual utterances (except for very few code-mixed utterances in the caregivers’ input; the children’s own monolingual data of course only contain monolingual data). Given that the traceback method can be considered a quite conservative approach (see e.g., Quick et al., 2021), it is actually quite remarkable that about 50% of all utterances can be successfully derived (in the case of Fion). As expected, the traceback success is much lower for the cross-corpus results, both when tracing the patterns detected in the code-mixed data to the input and when tracing them to the monolingual data. A mixed-effects logistic regression model using traceback success as the response variable, traceback type (within-corpus vs. cross-corpus) as the predictor variable, and the utterance as random effect shows that the difference is highly significant across both corpora, regardless of whether the child’s own monolingual data or the caregivers’ input is used as main corpus (**Table 1**).

Thus, our prediction that there are significant differences in traceback success between the within-corpus and the cross-corpus approach is confirmed. That being said, there is still much overlap between the results of both approaches, which indicates that a substantial number of patterns are shared between the two children. The individual differences between the two children rather become clear in another aspect of the results: Across the board, the traceback success for Silvie is much lower than for Fion. This also holds if we use each child’s entire dataset as test corpus, as shown in **Figure 3** (note that the overall traceback success is much higher if the monolingual utterances are taken into account). Again, the differences are highly significant according to a binomial mixed-effects regression model (**Table 1**).

The difference in traceback success between the two children is in line with the previous studies mentioned above, which have shown that Silvie’s language development is, overall, more advanced than Fion’s. As her utterances are longer and her grammar is more complex, the traceback method will inevitably produce more fails.

While the proportions of successful and failed tracebacks can be highly instructive, it can only be the first step of any traceback study. When investigating code-mixed utterances, it can be particularly revealing to take a closer look at the patterns detected by the traceback algorithm. **Table 2** shows the most frequently attested patterns for each child. Again, it has to be

<sup>3</sup>In future studies, we plan to enrich the corpus with part-of-speech tags to allow for a more fine-grained automatic analysis.



emphasized that our exploratory use of the traceback method claims no cognitive plausibility of the detected patterns but only serves as a proof-of-concept that the code-mixed utterances can, in principle, be accounted for using frame-and-slot patterns. Both children tend to combine *this* with German material—in the case of Fion, this pattern even accounts for no < 100 utterances. Apart from that, the patterns are relatively similar, and they substantiate the usage-based assumption that most instances of code-mixing can be accounted for with the help of simple frame-and-slot patterns. As a very rough tendency, the “frames” of *wh*-questions and simple assertive sentences (*ich bin kein* “I am no”) tend to come from German in both Fion’s and Silvie’s case, while

frequently attested content words like *fire*, *water*, *cheese* come from English.

Note that in all cases, not only are the fixed strings in the frame-and-slot pattern attested in the main corpus (here: the child’s own monolingual data) but also the fillers that occur in the open slots in the individual utterances. The fact that a large proportion of code-mixed utterances can be derived successfully with the help of the traceback method lends further support to the usage-based assumption that children’s early language use is strongly item-based, i.e., structured around concrete exemplars. This also becomes clear if we take a look at the data from a different but related perspective, by focusing on the

word pairs (bigrams) attested in the code-mixed utterances as shown in **Figure 4**. This figure shows, for each word (type) in the corpus, the immediately succeeding words attested in each child's utterances. The transparency of the arrows indicates the transition probability between words: Highly frequent bigrams

are indicated by a black arrow, while rare bigrams are indicated by a transparent (hence, gray-ish) arrow. For instance, in Fion's earliest data depicted in the upper-left panel of the plot, **und** "and" (highlighted with boldface) is strongly connected to *this*, as indicated by the thick black arrow. This means that the bigram **und this** occurs frequently in the data. The other word

**TABLE 1** | Coefficients of mixed-effects logistic regression models fit to the traceback data.

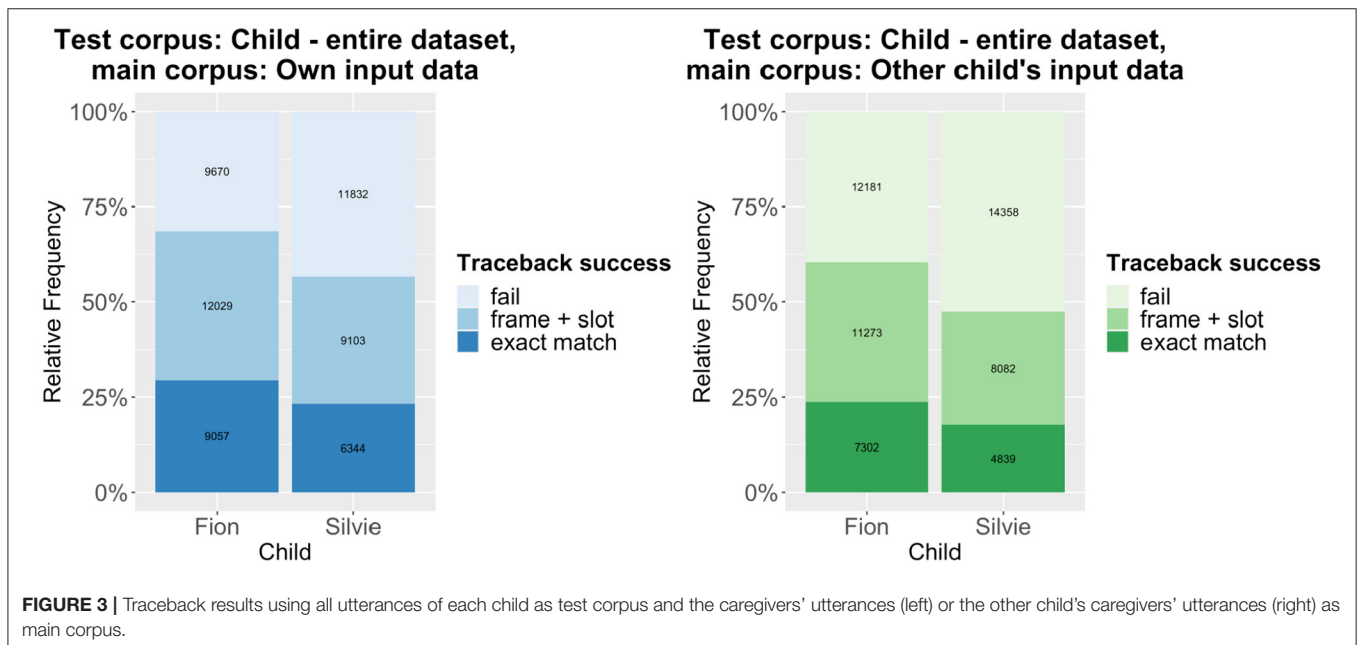
|   | Estimate    | Std. Error | z value      | Pr(> z )     |
|---|-------------|------------|--------------|--------------|
| <b>Code-mixed data → Monolingual data</b> |             |            |              |              |
| <b>Fion</b>                               |             |            |              |              |
| (Intercept)                               | -8.8        | 0.22       | -40          | 0***         |
| Type: cross-corpus                        | -7.6        | 0.27       | -28          | 5.20E-170*** |
| <b>Silvie</b>                             |             |            |              |              |
| (Intercept)                               | -9.2        | 0.21       | -43          | 0***         |
| Type: cross-corpus                        | -6.3        | 0.3        | -21          | 4.70E-99***  |
| <b>Code-mixed data → Input data</b>       |             |            |              |              |
| <b>Fion</b>                               |             |            |              |              |
| (Intercept)                               | -8.3        | 0.22       | -37          | 2.80E-305*** |
| Type: cross-corpus                        | -4.5        | 0.32       | -14          | 6.50E-45***  |
| <b>Silvie</b>                             |             |            |              |              |
| (Intercept)                               | -9.5        | 0.23       | -42          | 0***         |
| Type: cross-corpus                        | -6.2        | 0.46       | -13          | 1.00E-40***  |
| <b>All child utterances → Input data</b>  |             |            |              |              |
| <b>Fion</b>                               |             |            |              |              |
| (Intercept)                               | 1.57        | 0.051      | 3.076E + 01  | 1.0E-207***  |
| Type: cross-corpus                        | -1.20E + 00 | 0.034      | -3.600E + 01 | 7.5E-281***  |
| <b>Silvie</b>                             |             |            |              |              |
| (Intercept)                               | -0.18       | 0.048      | -3.70        | 2.0E-04***   |
| Type: cross-corpus                        | -1.30       | 0.035      | -36.00       | 1.6E-287***  |

\*\*\* $p < 0.001$ .

**TABLE 2** | Twenty most frequently attested patterns detected by the traceback method.

| Pattern             | Freq_Fion | Pattern                | Freq_Silvie |
|---------------------|-----------|------------------------|-------------|
| __ this             | 100       | __ this                | 24          |
| __ da               | 20        | __ this one            | 20          |
| <b>noch mehr</b> __ | 16        | <b>ein</b> __          | 14          |
| __ that             | 15        | __ cheese              | 11          |
| <b>ein</b> __       | 11        | <b>und</b> __          | 11          |
| <b>ich will</b> __  | 11        | <b>das ist</b> __      | 10          |
| __ you              | 11        | __ red                 | 9           |
| <b>und ein</b> __   | 10        | <b>nein</b> __         | 8           |
| __ cheese           | 9         | __ blue                | 8           |
| __ <b>feuer</b>     | 9         | <b>noch ein</b> __     | 8           |
| __ <b>nicht</b>     | 9         | __ green               | 8           |
| <b>nein ein</b> __  | 8         | <b>du bist</b> __      | 8           |
| __ milk             | 8         | <b>ich bin</b> __      | 8           |
| <b>das istis</b> __ | 7         | __ water               | 8           |
| <b>ich</b> __       | 7         | <b>ich bin kein</b> __ | 7           |
| this __             | 7         | __ <b>auch</b>         | 6           |
| <b>was ist</b> __   | 7         | <b>ja</b> __           | 6           |
| <b>wo eine</b> __   | 7         | __ cake                | 6           |
| <b>wo istis</b> __  | 7         | <b>der</b> __          | 6           |
| <b>wo kommt</b> __  | 7         | <b>du bist ein</b> __  | 6           |

German materials are indicated in bold.





We also found that the traceback success is significantly lower in the case of the cross-corpus approach, i.e., when tracing utterances from one corpus to the other. This is not surprising as each child has a different input situation from which they extract their individual linguistic knowledge. It is also clear that differences cannot be too large and that speakers need to converge on an inventory that overlaps in order to understand each other. However, there are also considerable individual differences that become clear when we look at the overall traceback success, which is lower for Silvie's data. All in all, Silvie's language is more complex, which also becomes clear in the bigram network that we have used to take a bird's-eye perspective on the children's code-mixing. Both the traceback study and the inspection of the bigram network, however, substantiate our hypothesis that the usage-based theory of language acquisition, according to which children's early utterances are organized around a limited set of concrete items (e.g., Tomasello, 2003), can account not only for monolingual, but also for multilingual language acquisition, and even for code-mixing.

However, the limitations of the traceback method, as discussed in e.g., Hartmann et al. (2021), should be kept in mind. Perhaps most importantly, the traceback method is largely limited to the detection of distributional patterns. As such, it presupposes a construction-centered approach to language (acquisition). As e.g., Wasserscheidt (2020: p. 61) points out, there is a long-standing debate between lexically-oriented and construction-based approaches. From a language acquisition point of view, Behrens (2007: p. 261) presents substantial empirical evidence in favor of "the construction as the primary conveyor of meaning." However, most usage-based theorists would readily acknowledge that lexical and constructional knowledge interact in language production and comprehension. As such, (syntactic) constructions alone are not enough to fully account for language acquisition. Drawing on an early precursor to the traceback method developed by Lieven et al. (1997), Vihman (1999) has argued that the role of semantic learning should not be underestimated: Her analysis of English-Estonian acquisition data suggest that predicate types and structures play a major role in the process of language learning. The traceback method, however, only identifies fixed strings and frame-and-slot patterns. It can at best provide indirect evidence regarding speakers' knowledge about the properties of individual lexical items. These caveats also apply to the exploratory study of bigrams. It would therefore be worthwhile to complement the inductive approach presented here with follow-up studies that take a different methodological perspective on the same data.

In addition, it should be kept in mind that our analysis was only exploratory, and follow-up studies should adopt a more fine-grained approach to the data. For one thing, adding a morphological annotation layer to the data could help to (semi-)automatically identify recurrent frame-and-slot patterns in a more reliable and psychologically plausible way. In addition, a systematic analysis of the failed tracebacks as conducted in previous traceback studies could add important insights (see e.g., Koch, 2019). Also, while we have only taken intra-utterance code-mixing into account, it would be interesting to investigate (emergent) individual differences in bilingual language acquisition against what is known from studies on

individual differences between adult speakers. For example, Street and Dabrowska (2010) have shown that individual differences are more likely for infrequently used constructions, whereas high-frequency syntactic patterns show less variability. Multilingual acquisition data can help us understand how such individual differences in language attainment come about and what role the linguistic input plays in this respect. Another desideratum would be to account for individual differences in bilingual speakers' attainment of the different languages they acquire—after all, it is well-known that bilingual speakers differ in their personal fluencies along various dimensions (see e.g., Edwards, 2013: pp. 11–14). The corpora analyzed in the present paper provide a rich source of data for approaching these and related questions in subsequent studies. In the long run, it would be desirable to extend the approach to other language pairs, which could prove insightful not only from a theoretical but also from a methodological perspective, as it could help to explore whether the traceback method can be fruitfully applied to pairs of typologically very different languages: In the case of German and English, the method works well as the structure of both languages is fairly similar, even though German has a slightly freer word order (see Koch, 2019: p. 212). But for another language, Italian, Miorelli (2017) has shown that the method meets some challenges, which are of course amplified when two languages are involved. The exploratory results presented in this paper can therefore only be a starting point for a broader theoretical and methodological discussion of how code-mixing should be modeled from a usage-based point of view.

## DATA AVAILABILITY STATEMENT

The data used for the present study are not yet publicly available but can be obtained upon request. The code used for data analysis as well as samples of the data are available at [https://hartmast.github.io/Quick\\_Hartmann\\_2021\\_Building\\_Blocks/](https://hartmast.github.io/Quick_Hartmann_2021_Building_Blocks/).

## 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 to participate in this study was provided by the participants' legal guardian/next of kin.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

## FUNDING

We acknowledge support from Leipzig University for Open Access Publishing.

## ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful comments on a previous version of this paper. The usual disclaimers apply.



## REFERENCES

- Ambridge, B., and Lieven, E. (2011). *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge: Cambridge University Press.
- Backus, A. (2021). "Usage-based approaches," in *The Routledge Handbook of Language Contact*, eds E. Adamou and Y. Matras (London: Routledge), 110–126.
- Behrens, H. (2007). "The acquisition of argument structure," in *Valency: Theoretical, Descriptive, and Cognitive Issues*, eds T. Herbst and K. Götz-Votteler (Berlin, New York, NY: Mouton de Gruyter), 193–214.
- Bernardini, P., and Schlyter, S. (2004). Growing syntactic structure and code-mixing in the weaker language: the Ivy hypothesis. *Biling. Lang. Cogn.* 7, 49–69. doi: 10.1017/S1366728904001270
- Bullock, B. E., and Toribio, A. J. (eds.). (2009). *The Cambridge Handbook of Linguistic Code-Switching*. Cambridge: Cambridge University Press.
- Cameron-Faulkner, T. H., Lieven, E., and Tomasello, M. (2003). A construction based analysis of child directed speech. *Cogn. Sci.* 27, 843–873. doi: 10.1207/s15516709cog2706\_2
- Cantone, K. (2007). *Code-Switching in Bilingual Children*. Dordrecht: Springer.
- Dabrowska, E., and Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cogn. Linguist.* 16, 437–474. doi: 10.1515/cog.2005.16.3.437
- Di Sciullo, A.-M., Muysken, P., and Singh, R. (1986). Government and code-switching. *J. Linguist.* 22, 1–24. doi: 10.1017/S0022226700010537
- Diessel, H. (2019). *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge: Cambridge University Press.
- Edwards, J. (2013). "Bilingualism and multilingualism: Some central concepts," in *The Handbook of Bilingualism and Multilingualism*, eds T. K. Bhatia and W. C. Ritchie (Malden, MA: Wiley-Blackwell), 5–25.
- Gardner-Chloros, P. (2009). *Code-Switching*. Cambridge: Cambridge University Press.
- Hartmann, S., Koch, N., and Quick, A. E. (2021). The traceback method in child language acquisition research: identifying patterns in early speech. *Langu. Cogn.* 13, 1–27. doi: 10.1017/langcog.2021.1
- Havron, N., and Arnon, I. (2021). Starting big: the effect of unit size on language learning in children and adults. *J. Child Lang.* 48, 244–260. doi: 10.1017/S0305000920000264
- Ibbotson, P. (2020). *What It Takes to Talk—Exploring Developmental Cognitive Linguistics*. Berlin: De Gruyter.
- Koch, N. (2019). *Schemata im Erstspracherwerb: eine Traceback-Studie für das Deutsche. (Linguistik, Impulse und Tendenzen 80)*. Berlin, Boston: De Gruyter.
- Koch, N., and Günther, K. (2021). Transfer phenomena in bilingual language acquisition: the case of caused-motion constructions. *Languages* 6:25. doi: 10.3390/languages6010025
- Koch, N., Hartmann, S.t., and Quick, A. E. (2020). The traceback method and the early constructicon: theoretical and methodological considerations. *Corpus Linguist. Linguist. Theory* doi: 10.1515/cllt-2020-0045. [Epub ahead of print].
- Lieven, E., Behrens, H., Speares, J., and Tomasello, M. (2003). Early syntactic creativity: a usage based approach. *J. Child Lang.* 30, 333–370. doi: 10.1017/S0305000903005592
- Lieven, E., Salomo, D., and Tomasello, M. (2009). Two-year-old children's production of multiword utterances: a usage-based analysis. *Cogn. Linguist.* 20, 481–508. doi: 10.1515/COGL.2009.022
- Lieven, E. V. M., Pine, J. M., and Baldwin, G. (1997). Lexically-based learning and early grammatical development. *J. Child Lang.* 24, 187–219. doi: 10.1017/S0305000996002930
- MacSwan, J. (2000). The architecture of the bilingual language faculty: evidence from codeswitching. *Biling. Lang. Cogn.* 3, 37–54. doi: 10.1017/S136672890000122
- Miorelli, L. (2017). *The Development of Morpho-Syntactic Competence in Italian-Speaking Children: a Usage-Based Approach*. Newcastle upon Tyne: Northumbria University. Available online at: <http://nrl.northumbria.ac.uk/id/eprint/39990>.
- Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.
- Myers-Scotton, C. (1997). *Dueling Languages: Grammatical Structure in Code Switching*. Oxford: Clarendon Press.
- Myers-Scotton, C., and Jake, J. L. (2015). "Cross-language asymmetries in code-switching patterns. Implications for bilingual language production," in *The Cambridge Handbook of Bilingual Processing (Cambridge Handbooks in Language and Linguistics)* (Cambridge: Cambridge University Press), 416–458.
- Pedersen, T. L. (2021). *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. Manual*. Available online at: <https://CRAN.R-project.org/package=ggraph> (accessed May 23, 2020).
- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español. *Linguistics* 18, 581–618.
- Quick, A. E., Backus, A., and Lieven, E. (2018). "Partially schematic constructions as engines of development: evidence from German-English bilingual acquisition," in *Cognitive Contact Linguistics*, eds E. Zenner, A. Backus, and E. Winter-Froemel (Berlin, Boston: De Gruyter), 279–304.
- Quick, A. E., Backus, A., and Lieven, E. (2021). Entrenchment effects in code-mixing: individual differences in German-English bilingual children. *Cogn. Linguist.* 32, 319–348. doi: 10.1515/cog-2020-0036
- Quick, A. E., Hartmann, S.t., Backus, A., and Lieven, E. (2019). Entrenchment and productivity: the role of input in the code-mixing of a German-English bilingual child. *Appl. Linguist. Rev.* doi: 10.1515/applirev-2019-0027. [Epub ahead of print].
- Street, J., and Dabrowska, E. (2010). More individual differences in language attainment: how much do adult native speakers of English know about passives and quantifiers? *Lingua* 120, 2080–2094. doi: 10.1016/j.lingua.2010.01.004
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge: Harvard University Press.
- Tomasello, M. (2009). "The usage-based theory of language acquisition," in *The Cambridge Handbook of Child Language*, ed E. L. Bavin (Cambridge: Cambridge University Press), 69–87.
- Vihman, M. M. (1999). The transition to grammar in a bilingual child: positional patterns, model learning, and relational words. *Int. J. Biling.* 2–3, 267–301. doi: 10.1177/13670069990030020801
- Vihman, V.-A. (2018). Language interaction in emergent grammars: morphology and word order in bilingual children's code-switching. *Languages* 3:40. doi: 10.3390/languages3040040
- Wasserscheidt, P. (2020). Explaining code switching. Matrix-language models vs. bilingual construction grammar. *Književni jezik* 31, 55–85. doi: 10.33669/KJ2020-31-04
- Yang, C. D. (2016). *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. Cambridge, MA: The MIT Press.

**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.

Copyright © 2021 Quick and Hartmann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.