frontiers in Computer Science

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fcomp.2021.659539

Speech Pauses and Pronominal Anaphors

Costanza Navarretta *

Centre for Language Technology, Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark

This paper addresses the usefulness of speech pauses for determining whether third person neuter gender singular pronouns refer to individual or abstract entities in Danish spoken language. The annotations of dyadic map task dialogues and spontaneous first encounters are analyzed and used in machine learning experiments act to automatically identify the anaphoric functions of pronouns and the type of abstract reference. The analysis of the data shows that abstract reference is more often performed by marked (stressed or demonstrative pronouns) than by unmarked personal pronouns in Danish speech as in English, and therefore previous studies of abstract reference in the former language are corrected. The data also show that silent and filled pauses precede significantly more often third person singular neuter gender pronouns when they refer to abstract entities than when they refer to individual entities. Since abstract entities are not the most salient ones and referring to them is cognitively more hard than referring to individual entities, pauses signal this complex processes. This is in line with perception studies, which connect pauses with the expression of abstract or complex concepts. We also found that unmarked pronouns referring to an entity type usually referred to by a marked pronoun are significantly more often preceded by a speech pause than marked pronouns with the same referent type. This indicates that speech pauses can also signal that the referent of a pronoun of a certain type is not the most expected one. Finally, language models were produced from the annotated map task and first encounter dialogues in order to train machine learning experiments to predict the function of third person neuter gender singular pronouns as a first step toward the identification of the anaphoric antecedents. The language models from the map task dialogues were also used for training classifiers to determine the referent type (speech act, event, fact or proposition) of abstract anaphors. In all cases, the best results were obtained by a multilayer perceptron with an F1-score between 0.52 and 0.67 for the three-class function prediction task and of 0.73 for the referential type prediction.

Keywords: speech pauses, abstract pronominal anaphora, individual pronominal anaphora, machine learning, corpus annotation, map task dialogues, first encounters

OPEN ACCESS

Edited by:
Anna Esposito, University of Campania Luigi Vanvitelli, Italy

Reviewed by:
Rubén San-Segundo, Polytechnic University of Madrid, Spain
Dubravko Culibrk, University of Novi Sad, Serbia

***Correspondence:**
Costanza Navarretta
costanza@hum.ku.dk

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *27 January 2021*
Accepted: *28 April 2021*
Published: *13 May 2021*

Citation:
Navarretta C (2021) Speech Pauses and Pronominal Anaphors. Front. Comput. Sci. 3:659539. doi: 10.3389/fcomp.2021.659539

1 INTRODUCTION

This paper addresses the role of speech pauses for determining what third person neuter gender singular pronouns, such as the English *it*, *this* and *that* refer to. The pronouns we address are pronominal intersentential anaphors, which can refer to concrete or abstract entities depending on their context. In example 1, the pronoun *it* refers to the glass on the table in the given communicative setting and has the nominal phrase *the glass* as antecedent. Instead in example 2, the pronoun refers to the generic event of cleaning the machines before and after use, and points back to the clause *Clean the machines before and after use*.

1. The glass is on the table. Please bring *it* to me. (antecedent: the glass).
2. Clean the machines before and after use. Always do *it* carefully. (antecedent: clean the machines before and after use.)

The other abstract pronouns in English are the demonstrative *this* and *that*, which often refer to complex abstract entities of different type. Pronominal anaphors that refer to individual entities as in 1 are called individual anaphors, while they are called abstract anaphors, discourse deictics (Webber, 1988), or situational referents (Fraurud, 1992) if they refer to an abstract entity as in 2. Third person neuter gender singular pronouns, *tpngs* pronouns henceforth, are quite common in all languages. Determining what they refer to, especially in the case of abstract anaphors, can be difficult for humans, and it is one of the most complex tasks in natural language processing. The task of finding the referent of referring expressions is called coreference resolution. In most natural language processing (NLP) systems, the task is simplified and resolution is redefined as finding the immediate *antecedents* of referring expressions. This process results in the annotation of co-referential chains of varying length as well as of expressions that are only mentioned once in discourse, the so called singletons. Finding the antecedents of intersentential pronominal anaphors is a sub-task of coreference resolution, called pronominal anaphora resolution. Most coreference systems only address expressions referring to individual entities, but more recently also events are dealt with e.g., Yang and Mitchell (2016), Barhom et al. (2019).

In this paper, we investigate the phenomenon of abstract and individual *tpngs* anaphors in Danish dialogues focusing on the importance of stress (accent) information on words and of speech pauses preceding the anaphors in order to determine whether they refer to individual or abstract entities or have other functions. Stress on words has especially been studied in terms of information structure, because words that are marked by a main accent are emphasized by the speaker and therefore become more salient to the addressee. Since salience is one of the main factors influencing the interpretation of pronominal anaphors, stress information is relevant in our context. Moreover, speech pauses have been interpreted as cognitive signals showing that the speaker is going to produce an abstract or complex concept Rochester (1973), Reynolds and Paivio (1968) or is going to refer to an object who has not the highest salience level e.g., (Gargiulo et al., 2019). More precisely, we analyze the relation between stress, speech pauses and *tpngs* pronouns in Danish annotated dialogues and then use this relation in language models on which we train classifiers to discriminate individual and abstract anaphors from other uses of these pronouns, and to predict the referent type of abstract anaphors. For the first classification task, the annotations of map task and first encounters dialogues are used, while for the second task we only use the annotations of the map task dialogues.

In **section 2**, studies about individual and abstract pronominal anaphora are presented, and the Danish *tpngs* pronouns are introduced. In **section 3**, relevant research on speech pauses is discussed. In the following **section 4**, the data used in this work is described. Moreover, the occurrences of abstract and individual *tpngs* pronouns in the two dialogue types and the role of stress and speech pauses for discriminating their uses are analyzed. In **section 5**, machine learning experiments are presented in which classifiers are trained on language models consisting of words alone, or enriched with accent information and speech pauses, in order to identify the pronouns' function and referent type. Finally, we discuss the results of our study in **section 6** and conclude in **section 7**.

2 INDIVIDUAL AND ABSTRACT PRONOMINAL REFERENCE

When we talk or write, we refer continuously to entities through nominal phrases. These can be substantives, with or without determiners and modifiers, or personal and demonstrative pronouns. Since *tpngs* pronouns can refer to either individual or abstract entities, it is first necessary to determine whether the referent is abstract or concrete in order to find their antecedent.

According to all researchers, who have presented cognitive models that account for the choice of referring expression by a speaker, this choice is based on assumptions the speaker makes about the varying status of entities in the addressee's mental state, see inter alia Prince (1981), Prince (1992), Gundel et al. (1993), Givón (1979), Ariel (1988), Ariel (2001). In all models, pronouns refer to the most salient or accessible entities. However, salience or accessibility is defined differently in the various models, even though all definitions are related. Prince (1981) and Prince (1992) looks at information structure and considers the most salient entities those that are oldest or known for longest time in discourse. Givón (1979) proposes to consider the entities that are most *topic prominent* at that point in discourse as being the most salient ones. According to (Gundel et al., 1993) instead, the most salient entities are those that are *in focus*. They also propose a scale of salient referring expressions, the *Givenness Hierarchy*. Finally, Ariel (2001) introduces the concept of *accessibility* level of entities, and she operationalizes it in terms of the distance of the referring expressions from their antecedents. In Ariel's model, the least marked and therefore easiest accessible referring expressions are zero pronouns, while the most marked expressions are full names with modifiers. Ariel's accessibility scale also distinguishes

between unstressed and stressed pronouns, the first being more accessible than stressed pronouns.

Kameyama (1998) studies the occurrences of unstressed and stressed versions of the same pronouns when they occur in the same position in discourse. She concludes that they have the same denotational range but indicate different preferred values. Because stressed pronouns signal a different presupposition than their unstressed counterparts, stressed pronouns take the complementary preference of their unstressed equivalents, when there are competing antecedents. Kameyama implements these findings in a system of preferences added to the Centering framework.

Webber (1988) finds that the English personal pronouns *it* does not often refer to abstract entities when the antecedents are clauses, because clauses are not easily accessible in discourse. The demonstrative pronouns *this* and *that* are instead used in these cases. Gundel et al. (1993) find the same pattern in their data and explain the preferred use of demonstrative pronouns with clausal antecedents in terms of their Givenness Hierarchy. Entities introduced in discourse by clauses are only activated in the hearer's cognitive status, being their referents facts, situation and propositions. On the contrary, eventualities, which are introduced in discourse by verbal or nominal phrases, often occur in a central syntactic position in the current or in the preceding utterance, and they are therefore often *in focus* and can be referred to by personal pronouns.

The preference for using demonstrative pronouns when referring to clauses or utterances as well as to discourse segments has been implemented in few rule-based resolution systems for English e.g., Eckert and Strube (2001), Byron (2002), Strube and Müller (2003), Müller (2007). An adaptation of Eckert and Strube (2001)'s algorithm to Danish was presented in Navarretta (2002) and Navarretta (2004).

The corpora annotated with abstract pronominal anaphora are only few (Kolhatkar et al., 2018), comprising the ARRAU corpus, which consists of both texts and transcriptions of dialogues (Poesio and Artstein, 2008). Part of the ARRAU corpus, the Wall Street Journal (WSJ) texts, has been used in the first benchmark for the resolution of *tpngs* anaphors in English (Marasovic et al., 2017). More precisely, Marasovic et al. (2017) run an LSTM-Siamese Net on the ARRAU annotations and report a precision score (s@1 which indicates that the correct antecedent span has been marked) of 29.06. Moreover, Poesio et al. (2018) describe how the ARRAU corpus was prepared to find the antecedents of abstract pronominal anaphors for a shared competition (CRAC, 2018), but no research group has addressed this task. The most recent work in which coreference resolution includes abstract pronouns is (Uryupina et al., 2020). The authors report an F1-score of 49.18 for the identification of non anaphoric expressions and singletons achieved by a LSTM. All morphological and syntactic features annotated in the data were included as well as all types of referring expressions in the ARRAU corpus.

Not only it is difficult to find the correct antecedents of abstract anaphors as shown in (Marasovic et al., 2017), it is also difficult to decide what the referent is (a task not addressed in anaphora resolution), because the anaphors often create their referent in the moment they point back to an antecedent Webber (1991), Fraurud (1992). Examples of how the same clausal antecedent of an abstract anaphor has different referents, depending on the context in which the anaphor is uttered, are in 3a–3c.

3. Peter was injured in a car accident
   a. When did *that* happen? (an event)
   b. *That* is not true. (a proposition)
   c. I did not know *that*. (a fact)

The three occurrences of the demonstrative *that* in the three utterances 3a–3c have the same antecedent, the utterance 3 *Peter was injured in a car accident*. However, in the example 3a, the referent is an event, in 3b it is a proposition and in 3b is a fact. Fraurud (1992) proposes that all abstract situations, defined by Vendler (1963) can be the referents of abstract anaphors. Asher (1993) builds a hierarchy of what he calls saturated abstract objects ordering abstract entities with respect to their degree of abstractness. The less abstract entities are eventualities, while the most abstract ones are proposition-like entities. Fact-like entities are in the middle. Also according to Asher, the entities which have the lowest degree of abstractness, can be referred to by personal pronouns as also noticed by Webber (1991), Gundel et al. (1993), while the most abstract entities are referred to by demonstrative pronouns.

The most common Danish *tpngs* pronoun is *det*, which corresponds to the three English pronouns *it*, *this* and *that*. In written language, *det* is ambiguous since it is not possible to determine whether it is a personal or a demonstrative pronoun. In spoken language, the unstressed *det* is the personal pronoun, while its stressed version *d,et*[1] is a demonstrative pronoun corresponding to both the English *this* and *that*. The demonstrative form can also occur as two words *det her* (this) and *det der* (that)[2]. Another Danish *tpngs* demonstrative pronoun is *dette* (this), which is only rarely used in spoken language.

The Danish *det* has many functions, and when it is used as anaphor, it can either refer to an individual or an abstract entity. In the first case, it can have as antecedents a nominal phrase referring to concrete entities in all genders and numbers, while in the second case it has other types of syntactic phrases as antecedents: verbal phrases, predicates in copula constructions, clauses and discourse segments (Navarretta, 2002). It has been found that Danish *tpngs* abstract pronouns occur in different contexts than the corresponding English pronouns and therefore theories and algorithms accounting for the use of these English pronouns cannot be directly applied to Danish data Navarretta (2002), Navarretta (2004). A Danish corpus annotated with abstract and individual *tpngs* anaphora was produced in the DAD project Navarretta and Olsen (2008). Our current study uses part of this corpus as well as the annotations of eight dyadic

[1]In this article, stress is marked with a comma preceding the stressed vowel.
[2]Allan et al. (1995) call the stressed versions of the pronoun *det* for *emphatic* pronouns.

first encounters from the Danish NOMCO corpus (Paggio and Navarretta, 2017) coded for this study.

3 RELEVANT STUDIES ON SPEECH PAUSES

Different types of speech pauses have been identified in discourse, the main types being silent pauses and filled pauses. Filled pauses are pauses accompanied by audible breaths or sighs, but can also occur with more or less lexicalized items, such as the English *ah* and *uh*. These items have been called fillers, discourse particle, and discourse markers.

Researchers have identified multiple functions of pauses involving both speech production and perception. For example, pauses have been found to be indicators that the speaker is planning what (s) he is going to say Maclay and Osgood (1959), Goldman-Eisler (1968), Chafe (1987), Hirschberg and Nakatani (1998), Shriberg (1994), and, they can limit syntactic phrases. Pauses can also signal that the speaker is searching for the correct lexical item (Krauss et al., 2000) and they are one of the multi-modal signals that contribute to regulate turn taking Duncan and Fiske (1977), Allwood (1988), Clark and Fox-Tree (2002).

Reynolds and Paivio (1968) found that silent and filled pauses occurred more frequently when students had to provide definitions of abstract objects than when they had to explain concrete objects. In line wit this study, Rochester (1973) investigated the frequency of hesitations in discourse and he concluded that their occurrences increase when speakers have to express something difficult or have to chose between more options. Finally, Esposito et al. (2002); Esposito and Esposito (2011) studied the occurrences of pauses and gestural pauses in more languages and propose that some of them have the function of introducing discourse new information either by speech or by gestures.

Navarretta (2007) analyzed the relation between types of Danish and Italian *tpngs* pronouns and their functions in an annotated abstract anaphora corpus, the DAD corpus (Navarretta and Olsen, 2008). She trained a support vector machine on the Danish annotations of texts, monologues and dialogues in order to identify these functions automatically (Navarretta, 2010) analyses the experiments focusing on the role of stress information in the spoken part of the data. Navarretta (2007)'s analysis of the data showed that there are numerous abstract and concrete pronominal occurrences in the data, and that both stressed and unstressed pronouns refer frequently to individual and abstract anaphors, confirming preceding studies that indicated that personal pronouns refer to abstract entities in Danish much more frequently than in English (Navarretta, 2004). In the classification experiments, run in WEKA, language models consisting of unigrams, bigrams and trigrams preceding and following the *tpngs* pronouns were used as training data and the best results were obtained when stress information was used and silent pauses were added to the language models consisting of different types of n-grams. The best results when classifying the functions of pronouns gave an F1 score of 0.51. The monologues contained only very few silent pauses, and the results of classification were extremely high (F1 score 0.982). These high results are due to the fact that the monologues consisted of the same texts read up by the various participants, and therefore train and test data were nearly the same with only small differences determined mostly by hesitations and placement of the stress by the different readers.

In a manually annotated spoken corpus, Roesiger and Riester (2015) find that information about the accent on words can contribute to coreference resolution. Moreover, Roesiger et al. (2017) test these findings on German data and conclude that pitch accents and phrasing improve coreference resolution. Gargiulo et al. (2019) study prosodic features and overt pronouns with individual antecedents in subject or object position in a production study involving Italian and Swedish speakers. They conclude that in both languages longer pauses precede inter-clausal pronouns when the antecedent is the less expected one. This is the subject for an overt pronoun in Italian, and the object in Swedish. Therefore, the function of pauses proposed by Gargiulo et al. (2019) is similar to the function of stress on the pronouns described by Kameyama (1998).

In this paper, we build upon the work in Navarretta (2007), Navarretta (2010) and re-use part of the data from those studies, as well as the newly annotated first encounters dialogues. The preceding studies are therefore extended in the following way: 1) both silent and filled pauses are considered in the present work, 2) a statistical analysis of differences of occurrences between individual and abstract pronouns is performed on the two types of dialogues, 3) anaphoric pronouns are studied together with the pauses which precede them in the maptask dialogues in order to determine whether they have specific uses, while Navarretta (2010) only added pauses to the training data as an extra prosodic feature without analyzing their possible functions, 4) we perform a number of classification experiments aimed to identify automatically three functions of *tpngs* Danish pronouns (individual, abstract entities or other functions) in the map task and first encounters dialogues, 5) we apply classifiers to the map task data in order to train classifiers to identify automatically the referent type of the abstract anaphors with verbal and clausal antecedents. In all experiments, we focus especially on the effect of pauses and stress information on classification.

4 THE DATA

4.1 The Annotated Corpora

The data used in this study consist of part of the DAD corpus and part of the NOMCO corpus.

The DAD corpus was collected and annotated under the project DAD, Det Abstrakt Det (the abstract it), funded by the Danish Research Councils. It consists of a collection of Danish and Italian texts and spoken data that were annotated with information about the antecedents and referents of *tpngs* pronouns. The annotations were made by two expert annotators (Navarretta and Olsen, 2008). In the present study, we only use the part of the DAD corpus consisting of the DanPASS dialogues (Grönnum, 2009). The DanPASS

dialogues are a Danish version of the map task dialogues described in (Anderson et al., 1991). In the map task dialogues, one participant guided the second participant in going through a map from the start point to an end point, the target. The two participants were sitting in two different rooms and could only speak together through head sets with microphones. The task was made more complex by the maps that the two participants worked with. The maps were similar, but they also had some differences, and the participants were not told about this.

The DanPASS corpus was collected, transcribed and phonetically annotated by phoneticians at the University of Copenhagen. The corpus consists of read texts, and map task dialogues. The participants were students and employees from the Department of General and Applied Linguistics, today part of the Department of Nordic Studies and Linguistics, at the same university. Stress, hesitations, pauses, tone of voice and other phonetic features were manually annotated by phoneticians (Grönnum, 2009) using PRAAT (Boersma and Weenink, 2009). The running words in the dialogue transcriptions are 52,145. In the DAD project, the transcriptions with pause and stress information were converted to XML and *tpngs* pronominal anaphora information were added using the PALinkA tool (Orăsan, 2003). The annotation scheme followed for coding anaphoric information was an extension of the MATE/GNOME annotation scheme for anaphora (Poesio, 2004). The extensions consisted mainly of elements and attributes added to the scheme in order to code non referential uses of the pronouns l and describe e.g., the abstract referent type. Most of the DAD data were annotated independently by the two annotators and then compared. Inter-coder agreement in terms of kappa scores Cohen (1960), Carletta (1996) was between 0.7 and 1 (Navarretta and Olsen, 2008), depending on the class and attribute. The data that were annotated by only one coder, was controlled by the second coder. In case of disagreement, the two annotators decided together which annotation to adopt. In difficult cases, linguist colleagues were consulted to choose the most probable annotation[3].

The Danish NOMCO corpus of first encounters consists of twelve spontaneous audio- and video-recorded dialogues between two young people who meet for the first time and talk freely for about fem minutes. The dialogues were collected and transcribed under the Nordic NOMCO project (Navarretta et al., 2012) at the University of Copenhagen. The transcriptions include pause and stress information annotated in the same format and system as in the DanPASS project, and the corpus has especially been used for studying multi-modal communication. Eight of the NOMCO encounters have been annotated with *tpngs* pronouns for this study. The *tpngs* pronoun annotations have been coded in an excel file following the annotation scheme and annotation manual produced by the DAD project. One dialogue was annotated by two coders independently, while the remaining dialogues were annotated by one coder. The intercoder agreement obtained for the annotations performed by two annotators in terms of kappa scores is between 0.58 and 1 depending on the category. These scores are lower than that obtained on the DAD corpus, which covered both texts and map task dialogues. The total of tokens comprised in these annotations are 13,300 and the total duration of the eight dialogues is of approx. 40 min.

4.2 The Annotations

The pronominal uses annotated in the data are the following:

- pleonastic as in *det sner* (it snows), *hun har det godt* (lit. she has it fine) (She is fine);
- cataphoric, the pronoun precedes the linguistic expression that is necessary to interpret it. *Det at Hanne ikke blev færdig med analysekapitlet til tiden, skabte problemer for hendes medstuderende.* (lit. It that Hanne did not finish the analysis chapter in time, gave problems to her fellow students) (The fact that Hanne did not finish the analysis chapter in due time, gave problems to her fellow students);
- deictic. The pronoun refers to an object in the physical context as in the following example: *Hvad er det der?*
- (What is that?)
- possibly accompanied by a pointing gesture to an object;
- individual anaphoric, the antecedent is a concrete entity: A: *Jeg har det forladte kloster lige i midten af kortet.* B: *Ja, det har jeg sådant set også. . .* (A: I have the abandoned closter precisely in the middle of the map B: I have it also, in a way . . .)
- individual vague anaphoric[4], the antecedent of the pronoun is a concrete entity that is implicit in discourse.
- abstract anaphoric, the antecedent is an abstract object: A: *okay der har jeg noget der hedder Den Blå Sø* B: *nej men det er jo helt forkert* (A: okay there I have something that is called The Blue Lake B: no but this is completely wrong)
- textual deictic (Lyons, 1977). "*Jeg elsker dig*" - ***Det*** *sagde han til hende for første gang, mens de snakkede.*("I love you" - He said that to her for the first time, while they talked together);
- abstract vague anaphoric-the abstract antecedent is implicit in the discourse;
- abandoned: the pronoun occurs in an unfinished utterance, which is then abandoned, and therefore it is not possible to infer the referent:
- *det er - han er gået*
- (it is - He is gone)

The type of referent of the abstract anaphors was also annotated in the map task dialogues as one of the extensions to the MATE/GNOME annotation scheme. The referent types that were identified are *eventuality*, *fact-like*, *fact-event*, *proposition-like*, *speech-act*. The type *fact-event* was assigned when the annotators found that the referent of the abstract anaphor was ambiguous and could be either a fact or an event.

[3]In some cases, when the annotators recognized that an anaphor could be interpreted in various way with the same probability, a class covering both readings was proposed. This was the case for the classification of the referent types.

[4]The use of the term *vague* in the list is taken from Eckert and Strube (2001).

TABLE 1 | Frequencies of pronominal uses in DanPass dialogues.

Pronoun	Indiv	IndVag	Abstr	AbstVag	Pleon	Cathaphor	Deic	Textdeict	Aband	Total
Unmarked	176	25	100	5	44	17	0	4	103	434
Marked	123	22	110	7	0	22	7	3	0	334
Total	299	47	219	12	44	39	7	7	103	768

4.3 Data Analysis

There are 768 *tpngs* pronouns in the DanPASS dialogues, and the occurrences of each type of pronoun are the following:

- *det*: 433
- *d,et*: 322
- *det her*: 1
- *d,et her*: 2
- *det h,er*: 1
- *det d,er*: 1
- *d,et h,er*: 3
- *det d,er*: 2
- *d,et der*: 1
- *det der*: 1
- *dette*: 1

The only occurrence of the demonstrative *dette* is an individual anaphor. When analyzing the demonstrative pronouns *det her* and *det der* and *dette*, we will consider them to be marked as the pronoun *d,et* independently from the presence and/or position of the stress. In **Table 1**, we show the functions of personal (non stressed/unmarked *det*) and demonstrative (marked) pronouns, that is *d,et*, *dette*, *det der*, and *det der*:

As expected, most occurrences of the pronouns have an anaphoric function, and the most common anaphoric use is that of referring to an individual entity (345 occurrences), followed by reference to abstract entities (223). 201 individual anaphors are unmarked, and 144 are marked. 105 abstract anaphors are unmarked, and 118 are marked. The difference between the use of unmarked and marked pronouns in reference to individual vs. abstract entities is statistically significant. The chi-square is 6.8076, the p-value is is 0.009077, with df = 1. The result is significant in the confidence interval $0.92 > p < 0.008$.

Thus, also in Danish spoken language there is a significant preference for referring to individual entities through unmarked pronouns and to abstract entities through marked ones. Textual deixis is performed by both marked and unmarked pronouns, and there are more marked cataphors than unmarked ones. Finally, abandoned pronouns are always unmarked in the DanPASS dialogues.

Out of the 345 occurrences of individual anaphors, 80 are preceded by a silent pause and 6 are preceded by a filled pause, that is 25% of the individual anaphors are preceded by a pause. 76 out of the 223 abstract anaphors are preceded by a silent pause and 9 of them by a filled pause, for a common total of 38% of their occurrences. Also in this case, the difference between the two types of anaphora is statistically significant. More precisely, considering the two types of pause together, the chi square statistic is 11.1973. The p-value is 0.000819 with degree of freedom = 1. The result is significant at $0.98 > p < 0.02$. Considering each type of pause separately, the chi square statistic is 11.9277. The p-value is 0.007637, with df = 3. The result in this case is significant in the confidence interval $0.0002 > p < 0.9998$. This shows that pauses preceding *tpngs* pronominal anaphors more frequently signal that the speaker is going to refer to an abstract entity than to an individual entity. As we have discussed previously, abstract reference is more complex and less expected than individual reference.

TABLE 2 | Abstract referent types and pronominal types in the DanPass dialogues.

Referent type	Pause + Unmark	Unmark	Pause + Marked	Marked	Total
Eventuality	7	19	9	31	66
Fact-event	1	1	1	0	3
Fact	15	13	17	32	77
Proposition	17	22	8	11	58
Speech-act	3	1	1	2	7
Total	43	56	36	76	211

In **Table 2**, the referential types of the various abstract pronouns in the DanPASS data are shown, distinguishing those that are preceded by pauses and those that are not.

The table shows that the most frequent referent type of abstract anaphors in these data is a fact-like entity, followed by eventualities and propositions. Surprisingly eventualities are more often referred to by a marked pronoun than by an unmarked pronoun, while propositions are more often referred to by unmarked pronouns than by marked ones. This seems to be contrary to what found for English by e.g., Webber (1988), Gundel et al. (1993) who notice that eventualities can be more easily referred by the personal pronoun *it* than other abstract entities since events are the most accessible abstract entities in discourse. It is interesting that there is a preference for pauses to precede unmarked pronouns more frequently when they refer to propositions and facts than when they precede eventualities. In this case, the chi-square statistic is 3.3403. The p-value is 0.067602 with df = 1. The confidence interval is $0.07 > p < 0.093$. This can be interpreted as a signal that pauses which precede a personal abstract pronoun in some cases signal that the referent of the anaphor is of a less expected type given its referent. This could be a similar function to that identified for other referential phenomena by (Gargiulo et al., 2019). However, the frequency of unmarked pronouns as referent of entities of higher abstractness degree also confirms the observation that there are language specific differences with respect to individual and

TABLE 3 | Frequencies of pronominal uses in the NOMCO encounters.

Pronoun	Indiv	IndVag	Abstr	AbstVag	Pleon	Cathaphor	Deic	Textdeict	Aband	Total
Unmarked	172	16	165	14	97	75	0	2	54	594
Marked	22	6	61	8	0	8	12	3	22	142
Total	194	21	226	22	97	81	12	5	76	736

abstract reference between Danish and English as discussed in Navarretta (2002), Navarretta (2004).

There are 736 *tpngs* pronouns in the annotated first encounters. The pronouns are distributed into the following types:

- *det:* 594
- *d,et:* 127
- *det her:* 1
- *d,et her:* 1
- *det der:* 4
- *det d'er:* 1
- *d,et der:* 7
- *d,et d,er:* 1

There are 594 unmarked and 142 *tpngs* pronouns in these dialogues. It is interesting to note that *tpngs* pronouns are much more frequent in the spontaneous first encounters than in the map task dialogues, since their relative frequency in the former corpus is 0.055, while it is 0.015 in the latter.

In **Table 3** are shown the different types of pronominal functions of the *tpngs* pronouns in the first encounters. Unmarked *tpngs* pronouns are more frequent in these dialogues than in the DanPASS dialogues, while individual *tpngs* anaphors are more frequent in the map task dialogues than in the first encounters. Finally, abstract *tpngs* anaphors are more frequent in the first encounters than in the map task dialogues.

The fact that individual anaphors are more frequent in the map task dialogues is not surprising given that the speakers often refer to individual objects on their maps. The higher frequency of abstract anaphors in the first encounters can explain the lower intercoder agreement obtained for these data, since individual anaphors are easier to annotate. In the first encounters, the unmarked pronoun is more often used as individual than as abstract anaphor, while the opposite holds for marked pronouns. The difference is also in these dialogues significant. In fact, the chi-square statistic is 15.417. The p-value is 0.000086 with df = 1 and the confidence interval holds for $0.0001 < p > 0.9999$. This difference confirms again that also in spoken Danish as in English in general there is a preference for unmarked pronouns to refer to individual objects, and for marked pronouns to refer to abstract pronouns. However, these data also confirm that unmarked pronouns are also used as abstract anaphors in Danish more frequently than in English. In fact, the pronoun *it* has been found to be abstract pronoun in less than one thir of its occurrences in e.g., Webber (1988), Gundel et al. (1993), Poesio and Artstein (2008). The analysis of pauses preceding *tpngs* pronoun types will be performed in future.

5 IDENTIFYING PRONOMINAL ANAPHORIC FUNCTIONS AND ABSTRACT REFERENT TYPES

In our classification experiments, we address the automatic classification of the individual and abstract functions of *tpngs* pronouns in the Danish map task and first encounters dialogues. Differing from the experiments presented by Navarretta (2007), Navarretta (2010), we include filled pauses to the annotations and focus on the classification of individual and abstract anaphors vs. all other uses of the pronouns. Moreover, a different classification strategy and more classifiers are applied.

In these experiments, the five categories cataphoric, pleonastic, deictic, textual deictic and abandoned have been collapsed in one class *other*, since our main aims are 1) to determine whether the information of stress and pauses can help disambiguating individual and abstract anaphors as well as distinguish them from other uses, and 2) to find the best data sets and classifiers for this task.

The DanPASS data sets contain 345 pronouns referring to an individual entity (the individual and individual vague pronouns), 223 pronouns referring to an abstract entity (the abstract and abstract vague pronouns), and 200 pronouns classified as *other*. The NOMCO data sets contain 216 individual pronouns (also in this case the pronouns classified as individual and individual vague), 248 abstract pronouns (those classified as abstract and abstract vague) and 272 pronouns classified as *other*. The supervised machine learning experiments were run in python 3.7 with the numpy and scikit-learn packages. The classifiers that were tested are K Neighbors Classifier (KNC), Multinomial Naive Bayes (MNB), Multilayer Perceptron (MP), Support Vector Machine (SVM), and Logistic Regression (LR).

The multilayer perceptron was run with the following hyper-parameters: the *adam* solver with the *tahn* activation, two layers of size 3 and 1, 5,000 iterations, adaptive learning rate and alpha = 0.001 for the DanPASS data. The parameters used for the NOMCO corpus are the *sgd* solver with the *tahn* activation, four layers of sizes 8, 5, 5, and 1, constant learning rate and alpha = 0.001. The optimal parameters were found applying the GridSearchcv method on a variety of hyper-parameters including two solvers, two learning rates and activation layers as well as different number of layers. The GridSearchcv was run on 20% of the data, then a different 20% of the data was used for testing, and finally, ten-fold cross-validation for the best performing classifier and data set were run to control that the results from testing are not due to overfitting.

The language models in the training data were the ones that gave the best results on the task of classifying all pronominal functions in (Navarretta, 2007). The models consist of six-grams,

TABLE 4 | Examples of various data sets.

Data set	Token1	Token2	Token3	Token4	Token5	Token6
Word	jeg	men	det	er	jo	Ikke
word&stress	jACCeg	men	det	er	jo	iACCkke
word&pause	FILPAUSE	men	det	er	jo	Ikke
All	FILPAUSE	men	Det	er	jo	iACCkke

TABLE 5 | Results of classifiers predicting pronominal type in the DanPASS dialogues.

Classifier	Dataset	P	R	F1
Majority		0.152	0.39	0.22
Random		0.322	0.312	0.314
KNC	Words	0.546	0.558	0.541
	words&accent	0.57	0.571	0.572
	words&pause	0.514	0.523	0.51
	All	0.585	0.591	0.582
MNB	Word	0.647	0.623	0.62
	word&accent	0.683	0.81	0.675
	word&pause	0.664	0.649	0.649
	All	0.698	0.681	0.682
SVM	Words	0.738	0.662	0.651
	word&accent	0.7	0.636	0.636
	word&pause	0.714	0.662	0.628
	All	0.71	0.623	0.6
MP	Word	0.638	0.617	0.619
	word&accent	0.645	0.643	0.643
	word&pause	0.65	0.63	0.64
	All	0.69	0.681	0.684
LR	Words	0.663	0.643	0.64
	word&accent	0.647	0.636	0.637
	word&pause	0.67	0.662	0.663
	All	0.685	0.67	0.67

TABLE 6 | Results of classifiers predicting pronominal type in NOMCO dialogues.

Classifier	Dataset	P	R	F1
Majority		0.114	0.339	0.171
Random		0.333	0.331	0.331
KNC	Words	0.466	0.466	0.466
	words&accent	0.46	0.46	0.46
	words&pause	0.393	0.4	0.395
	All	0.44	0.44	0.44
MNB	Word	0.523	0.51	0.5
	word&accent	0.49	0.472	0.471
	word&pause	0.463	0.446	0.433
	All	0.51	0.5	0.5
SVM	Words	0.49	0.48	0.49
	word&accent	0.51	0.49	0.49
	word&pause	0.478	0.453	0.423
	All	0.542	0.52	0.511
MP	Word	0.314	0.432	0.364
	word&accent	0.47	0.46	0.46
	word&pause	0.5	0.473	0.474
	All	0.524	0.53	0.521
LR	Words	0.51	05	0.493
	word&accent	0.472	0.46	0.46
	word&pause	0.47	0.46	0.453
	All	0.493	0.49	0.486

in which the *tpngs* pronoun whose function must be predicted, is the third speech token. A speech token can be a word, an hesitation or a pause. The context of each pronoun consists therefore of two speech tokens preceding the pronoun and three speech tokens following it. The data set types used in training are the following: 1) only word tokens, 2) word tokens with stress information, 3) speech tokens consisting of words, hesitation, silent or filled pauses, 4) all features, that is word tokens with eventual stress, hesitation or pause tokens. In **Table 4** a line from each data set is shown in order to illustrate the six-grams language models on which the pronouns were trained. In the data, stress is indicated by the string *ACC* preceding the stressed vowel e.g., *dACCet* is the marked pronoun *det*, while silent pauses are indicated as SILPAUSE and filled pauses are joined in the class FILPAUSE.

Differing from the experiments in Navarretta (2007), Navarretta (2010), stress information on the pronouns is kept in all data sets, since the usefulness of stress information on the pronouns has been demonstrated in other studies. This change allows us to compare exclusively the impact of the contextual information of the pronouns on classification.

In all the experiments, the baselines are a majority classifier, which chooses the most frequent class and a random classifier, which assigns a class randomly taking account of the frequency of the classes. The most frequent class in the map task dialogues is *individual*, while it is *other* in the first encounters.

The results of classification on the DanPASS data sets is in **Table 5**, and the results of classification on the NOMCO data sets are in **Table 6**. Precision (P), Recall (R) and weighted F1-score (FI) for each classifier and each data set is shown in the table.

All algorithms perform significantly better than both the majority and random classifiers, but their performance varies from data set to data set. The best results in both map task and first encounters dialogues were obtained by the Multilayer Perceptron trained on the data set including all prosodic features. On the DanPASS data the F1-score is 0.684. The second best performing classifier, on this data, is the Multinomial Naive Bayes Classifier which also performs well on the other data sets. The F1-score by the Multilayer Perceptron trained on the data with all features improves by 0.37 the results of the random classifier while the improvement with respect to the majority baseline is of 0.464. It must be noted that introducing information on pauses alone does not improve classification with respect to data where stress information or only words are used in many cases because the tokens consisting of words and stressed words in the context of pronominal anaphors are more discriminative than the two tokens *filled* or *silent* pause.

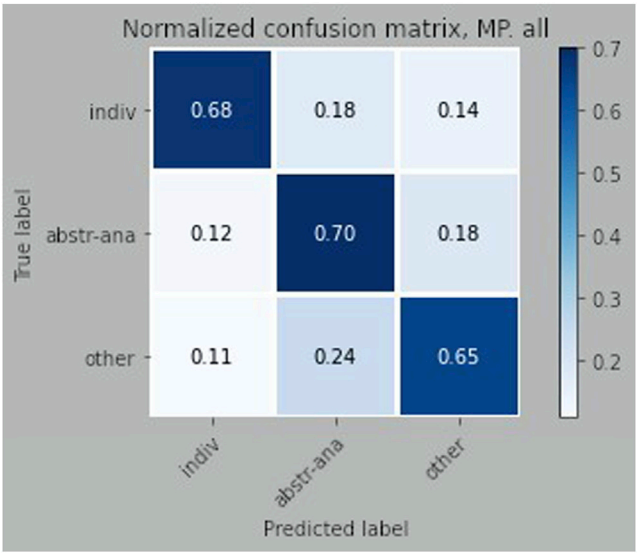

FIGURE 1 | Normalized confusion matrix for the DanPASS dialogues.

However, when both stress information and pauses are used, the best results are obtained. The normalized confusion matrix returned by the Multilayer Perceptron trained on the DanPASS data with stress and pause information is in **Figure 1**. The confusion matrix shows that the class that is predicted correctly more often is that of abstract reference. This is interesting since abstract anaphors are not the most frequent category in the data, and they are difficult to identify without looking at the contextual content Webber (1988), Fraurud (1992), Eckert and Strube (2001). For this reason, abstract anaphors are often excluded from coreference resolution systems. Moreover, the classifier also succeeded in distinguishing abstract and individual anaphors from other types of pronominal uses of *tpngs* pronouns. The averaged results of the ten-fold cross validation gives an F1-score of 0.641, showing that the performance falls only slightly.

The most frequently misclassified functions of pronouns are *other* uses that are classified as *abstract*. A nearer analysis of the wrong classified cases shows that especially cataphoric, and in less degree, abandoned uses of pronouns are misclassified as abstract, if the context included in the language model is not sufficient to discriminate the different uses. Similarly the misclassification between abstract and individual anaphors is mostly due to ambiguous cases in which only the larger context of the dialogue can, in most cases, lead to the correct interpretation of the pronouns. One example is the utterance *det var godt* (it was fine/good) that can both refer to an individual object, in same case independently from the gender of the antecedent, and an abstract entity depending on the context. In some cases, utterances of this type are ambiguous also for humans, and some utterances can have both readings. Only the preferred interpretation was annotated in the data.

Also in the case of the NOMCO data, all classifiers perform better than the two baselines. The best results are obtained by the Multilayer Perceptron with an F1 score of 0.521. This result improves with 0.35 the majority baseline, and of 0.19 the random baseline. One of the reasons for the lower results obtained on the spontaneous dialogues is that the individual anaphors that are the easiest class to identify are less frequent than the abstract anaphors, while the most frequently occurring pronominal type is the most ambiguous one, the unstressed *det*, which can have all the three functions with nearly the same frequency. The second best performed classifiers is also here the Multinomial Naive Bayes Classifier, which performs best when trained on word token six-grams and on all types of tokens. The averaged ten-fold cross validation gave an higher F1 score = 0.534.

The normalized confusion matrix returned by the Multilayer Perceptron with stress and pause information is in **Figure 2**. The confusion matrix shows that the class that is in most cases classified correctly is *other*. Abstract anaphors and individual anaphors are identified correctly with the same frequency, over 50% of the cases. The classes that are more often confused are *individual* and *other*. Looking at the misinterpreted occurrences, we found that it is especially cataphoric and abandoned uses of the unstressed pronoun *det*, which are confused with its individual uses and vice versa. This happens especially when the pronominal context contains a sequel of so called clausal adverbials, which in Danish precede or follow the finite verb in a clause depending on whether the clause is main or subordinate. The presence of these adverbials results in identical contexts for many types of pronominal uses. Examples of these adverbs in Danish are *da* (surely), *jo* (certainly), *vel* (presumably), *aldrig* (never), *ikke* (not). They can be combined in different ways and are frequent in especially spoken language. The presence of clausal adverbials in the six-grams language models is also one of the reasons behind some of the errors confounding abstract and individual anaphors. Moreover, cases in which only the larger context of the dialogue can indicates the correct interpretation of the pronoun, were found also in these data.

The last classification experiment aimed to determine the type of referent of the abstract anaphors in the DanPASS data. This information has not been annotated yet in the NOMCO data. The

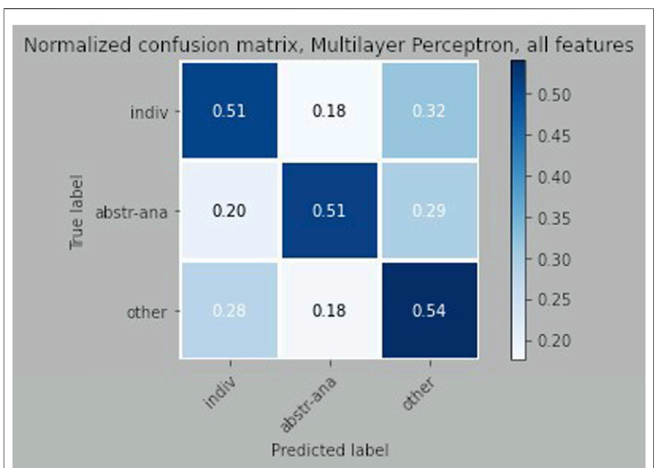

FIGURE 2 | Normalized confusion matrix for the NOMCO dialogues.

TABLE 7 | Results of classifiers predicting the referent type of abstract anaphors in the DanPASS dialogues.

Classifier	Dataset	P	R	F1
Majority		0.175	0.419	0.247
Random		0.274	0.209	0.23
KNC	Words	0.57	0.581	0.564
	words&accent	0.597	0.605	0.596
	words&pause	0.638	0.605	0.6
	All	0.639	0.628	0.0.617
MNB	Word	0.637	0.628	0.616
	word&accent	0.672	0.651	0.644
	word&pause	0.672	0.674	0.668
	All	0.71	0.698	0.698
SVM	Words	0.602	0.604	0.6
	word&accent	0.693	0.674	0.671
	word&pause	0.678	0.674	0.67
	All	0.707	0.698	0.67
MP	Word	0.68	0.7	0.69
	word&accent	0.638	0.651	0.644
	word&pause	0.622	0.651	0.631
	All	0.734	0.744	0.737
LR	Words	0.6	0.674	0.665
	word&accent	0.692	0.698	0.69
	word&pause	0.684	0.698	0.69
	All	0.733	0.744	0.736

data consists of 211 six-grams as in the preceding experiments, but only abstract pronouns have been held and information about their referent type is added as the class to be predicted. The ambiguous class "event-fact" is treated as an eventuality, thus the four classes to be predicted are *eventuality*, *fact-like*, *proposition* and *speech-act*. The best data set from the preceding experiment, that is the language model which contains all prosodic information was used. In **Table 7**, we present the results of the two baselines and the two best performing algorithms, which on this task and data are the K Neighbors Classifier (KNC) and the Multilayer Perceptron run with the following hyper-parameters: The *sgd* solver with the *relu* activation, two layers of size 6, and 1, 6,000 iterations, adaptive learning rate and alpha = 0.0001. Also in this table the results are presented in the form of weighted Precision, Recall and F1 score.

The best results were obtained again by the Multilayer Perceptron. The F1 score was 0.737, which is an improvement of 0.49 respect to the best performing baseline, in this case the majority classifier, and of 0.507 respect to the random classifier. Given that it is often difficult to determine the referent type also for humans, these results are very positive. In this case, most errors were confusing eventuality and fact readings of the pronoun. The F1 score from ten-fold cross validation is 0.71.

We also repeated the classification experiment, which gave the best results in Navarretta (2007) and that it is discussed in Navarretta (2010) in order to test whether the use of filled pauses decreases or increases the classification results. The original experiment was run in the WEKA SMO classifier (a support vector machine) with ten-fold cross validation and the aim was to classify all nine functions of the *tpngs* pronouns from the DanPASS dialogues. In the experiment, the effect of pauses was only measured together with stress information added to words, and the F1 score reported in Navarretta (2007) is 0.518. We run the scikit-learn SVM classifier with a linear kernel on the extended data set and obtained and F1-measure of 0.554. Running ten-fold cross-validation gave an F1-score of 0.549. These results show that information about filled pauses improves classification, even if there are few of them.

6 DISCUSSION

The analysis of the dialogue annotations show that even if personal and demonstrative *tpngs* pronouns are used frequently in the dialogues with both an individual and abstract anaphoric function, there is a preference for the marked pronouns to have an abstract antecedent and for the unmarked personal pronoun *det* to have individual antecedents. The preference is stronger in the first encounters than in the map task dialogues. Thus, the observation made by Navarretta (2004), Navarretta (2007) that in Danish the personal pronoun *det* is the preferred anaphor in both individual and abstract reference only holds in written language, where it is not possible to distinguish between unmarked and marked occurrences of the pronoun.

The analysis of the DanPASS dialogues also shows that both silent and filled pauses precede more frequently abstract anaphors than individual anaphors. This indicates that speech pauses can signal that the speaker is going to utter a difficult concept, since according to all theories of reference *tpngs* pronouns with abstract antecedents are less salient/accessible to the addressee than pronouns with individual antecedents, and they are also more difficult to express for the speaker. This work is therefore in line with perception studies that found that speakers use more frequently pauses when they have to utter or define abstract concepts than concrete ones Rochester (1973), Reynolds and Paivio (1968).

We also found that unmarked pronouns are the most common anaphors with a proposition referent even if propositions are the most abstract types of entities according to e.g., (Asher, 1993). On the other hand, in the DanPASS dialogues, marked anaphors are the most frequently occurring pronouns with referents of the eventuality type, which according to researchers investigating reference in English e.g., Webber (1988), Gundel et al. (1993), Asher (1993) are the less abstract type of abstract entity, and they are therefore often referred to by personal pronouns. The analysis of pauses preceding the abstract anaphors in our data also shows that personal pronouns with a fact or proposition referent are more often preceded by pauses than when they refer to eventualities. We propose that the presence of a pause, in these cases, might mark that the referent of the pronoun is not the most expected one. In this cases, pauses have a similar function as that observed by Gargiulo et al. (2019) on Italian and Swedish overt individual pronouns with subject or object antecedents. However, since the data are of limited size and we do not have other corpora to compare our data with, this supposition should be tested in more dialogues. Moreover, we did not notice a similar use of pauses when marked pronouns referred to eventualities. In general, with respect to reference, it is only possible to study preferred uses since other factors than salience

can move a speaker to use one form of reference instead of others, some of these being variation and personal preferences. In the future, we will also analyze the use of pauses in the first encounters.

The machine learning experiments aimed to determine to what extent individual, abstract anaphors and other uses of Danish *tpngs* pronouns can be predicted automatically training classifiers on language models of speech tokens were strongly inspired by the experiments performed by Navarretta (2007) on the same data. For example, the best performing language models in those experiments, six-grams, were used in the present work.

The experiments were also repeated on new annotated data, part of the spontaneous Danish NOMCO first encounters corpus. The results of our experiments show that a semi-automatically tuned Multilayer Perceptron can identify the correct function of *tpngs* pronouns in more than two thirds of their occurrences, with a weighted F1-score of 0.67 on the DanPASS data, and on 0.52 of the cases on the first encounters. Running ten-fold cross validation on the same data sets gave similar results. These are both significant improvements of the F1 score achieved by the random and majority classifiers. Interestingly, the class which is identified most correctly in the map task dialogues is that of abstract anaphors, even if they are not the most frequently occurring type in these data.

In the first encounters, the class that is identified correctly in more cases is the *other* class, which is the most frequent one. However, the confusion matrix from this experiment indicates that also abstract and individual uses of *tpngs* pronouns are correctly identified in over 50% of the cases. The analysis of erroneous classified entities shows that in the map task dialogues the classes most often confused are *other* and abstract anaphors. In particular cataphoric and abandoned pronouns are often confused with abstract anaphors and vice versa when the context included in the used language model is not large enough. In the first encounters, abstract anaphors and cataphoric or abandoned pronouns are the classes that are most often confused by the classifiers. Finally, insufficient contextual size was the cause of many errors classifying abstract anaphors as individual ones and vice versa.

The obtained results indicate that using information on speech pauses and stress on words can contribute significantly to the task of distinguishing individual, abstract pronominal anaphors, and other pronominal functions automatically. The automatic classification could replace or support rule-based discrimination of individual and abstract anaphors in anaphora resolution systems.

The fact that introducing information on pauses alone does not improve classification with respect to data where stress information or only words are used is due to the fact that tokens consisting of words and stressed words surrounding pronominal anaphors are more discriminative than the two tokens *filled* or *silent* pause. However, when pauses are added to words with stress information the classification results improve for most of the classifiers. It most also been noticed that running ten-fold cross classification with the Multilayer Perceptron gives a fall of approx. 0.04 in performance with respect to when we run the experiments training the data on 80% of the data, but they are still good indicating that the results are quite reliable on these data at least.

The best measure we have in order to compare our classification results with the current state of art is given by the results reported by Marasovic et al. (2017) for the resolution of English *tpngs* pronouns in the WSJ part of the ARRAU corpus. The results were obtained by LSTM trained on the many morphological and syntactic features in the corpus annotations. The precision of the machine learning algorithm is reported to be 29.01. They also report that the algorithm proposed the correct antecedent as the fourth ranked candidate in the antecedent candidate list in 63.55 of the cases. In these candidate list were often both nominal and verbal phrases, as well as clauses. Our results are not directly comparable, since Marasovic et al. (2017) not only identify, but also resolve English abstract and individual *tpngs* pronouns in approx. one third of their occurrences. However, the difference in performance on the resolution of *tpngs* pronouns compared to that obtained by coreference systems in general, shows clearly how difficult the present task is. Moreover, our results are interesting for different reasons. First of all, the Danish unmarked personal pronoun is used much more often than in English to refer to abstract entities when we compare our data with what has been reported for English e.g., Webber (1988), Gundel et al. (1993), Poesio and Artstein (2008). Therefore, abstract and individual anaphors in Danish are harder to discriminate on the basis of the pronominal type than in English. Secondly, our data are dialogues, which are notoriously more difficult to process than texts, while Marasovic et al. (2017) address newspapers. Finally, we have not relied on morphological and syntactic features as anaphor resolution systems do by choice, since our aim has been to investigate the importance of pauses and stress for distinguish between different functions of the Danish *tpngs* pronouns.

The six-grams language models were also used in order to identify the referent type of the abstract anaphors in the DanPASS data. This task has not been attempted on other languages, and it is also difficult for humans as discussed in e.g., Webber (1991), Fraurud (1992). We achieved good results, F1 score = 0.737, but we only run the classifiers on abstract anaphors. However, knowing the type of referent is useful to identify the correct antecedent of abstract anaphors, and this information is not present in the English corpora including abstract pronominal anaphora annotations.

7 CONCLUSIONS AND FUTURE STUDIES

In the paper, we presented a study of individual and abstract anaphoric reference and the role of speech pauses that precede them in two annotated types of Danish dyadic dialogues: map task dialogues and first encounters. The study has given new insight into the phenomenon of pronominal individual and abstract reference in Danish spoken language, pointing out the role of speech pauses in the task of determining whether *tpsng* pronouns are used anaphorically or not and, in the case they are

anaphors, whether they refer to an individual or an abstract entity.

This work revises preceding studies that concluded that the most common abstract pronoun in Danish is the unmarked pronoun *det* (Navarretta, 2010), since this is only valid for Danish texts. In both types of dialogues, in fact, there is a statistically significant preference for referring to individual entities with the unmarked personal pronoun *det*, while the stressed *d'et* and the demonstrative pronouns *det her* and *det der* have more often abstract referents as it is also the case for the corresponding pronouns in English. Secondly, we found that silent and filled pauses precede much more frequently abstract pronominal anaphors than individual pronominal anaphors, confirming preceding studies that showed that people produce more hesitations and silent pauses when they are going to utter an abstract concept than a concrete one (Rochester, 1973). Thirdly, on the basis of the analysis of speech pauses preceding abstract pronominal amphora with different types of referents, we propose tentatively that speech pauses signal that the referent of an unmarked anaphor is of a type less expected given the pronominal type.

Finally, our machine learning experiments confirm the fact that information about occurrences of speech pauses and stress information on words can classify individual anaphors, abstract anaphors or other functions of Danish *tpsng* pronouns in more than two thirds of their occurrences in the map task dialogues and in more than half cases in the first encounters. Adding information on filled pauses also helps classifying all uses of *tpsng* pronouns improving the experiments in (Navarretta, 2007).

Since the proposals in this paper are based on the analysis of one types of dialogue, the use of pronominal anaphors in spoken data and the function of pauses with respect to individual and abstract reference should be investigated in more spontaneous dialogues and in different domains. In the future, it should also be tested whether silence pauses can signal the occurrences of abstract anaphors or that the pronoun they precede has a referent of a more abstract type than that pre-announced by the pronominal type in other languages.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The DanPASS dialogues are publicly available, the DAD data and the NOMCO annotations can be obtained contacting the author. Requests to access these datasets should be directed to DanPASS, https://danpass.hum.ku.dk/, DAD and NOMCO annotations to costanza@hum.ku.dk.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Allan, R., Holmes, P., and Lundskær-Nielsen, T. (1995). *Danish - A Comprehensive Grammar*. London: Routledge.

Allwood, J. (1988). "The Structure of Dialog," in *Structure of Multimodal Dialog II*. Editors M. M. Taylor, F. Neél, and D. G. Bouwhuis (Amsterdam: John Benjamins), 3–24.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The Hcrc Map Task Corpus. *Lang. Speech* 34, 351–366.

Ariel, M. (2001). "Accessibility Theory: An Overview," in *Text Representation, Human Cognitive Processing Series*. Editors T. Sanders, J. Schliperoord, and W. Spooren (John Benjamins), 29–87.

Ariel, M. (1988). Referring and Accessibility. *J. Linguistics* 24, 65–87.

Asher, N. (1993). "Reference to Abstract Objects in Discourse," in *Studies in Linguistics and Philosophy* (Dordrecht, Netherlands: Kluwer Academic Publishers), Vol. 50.

Barhom, S., Shwartz, V., Eirew, A., Bugert, M., Reimers, N., and Dagan, I. (2019). "Revisiting Joint Modeling of Cross-Document Entity and Event Coreference Resolution," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Florence, Italy: Association for Computational Linguistics), 4179–4189.

Boersma, P., and Weenink, D. (2009). Praat: Doing Phonetics by Computer. Available at: http://www.praat.org/ (Retrieved May 1, 2009).

Byron, D. K. (2002). "Resolving Pronominal Reference to Abstract Entities," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02) (Philadelphia, PA: Association of Computational Linguistics), 80–87.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistics. *Comput. Linguist.* 22, 249–254.

Chafe, W. (1987). "Cognitive Constraint on Information Flow," in *Coherence and Grounding in Discourse*. Editor R. R. Tomlin (Amsterdam: John Benjamins), 20–51.

Clark, H. H., and Fox-Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition* 84, 73–111. doi:10.1016/s0010-0277(02)00017-3

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* 20, 37–46. doi:10.1177/001316446002000104

Duncan, S., and Fiske, D. (1977). *Face-to-Face Interaction*. Hillsdale, NJ: Erlbaum.

Eckert, M., and Strube, M. (2001). Dialogue Acts, Synchronising Units and Anaphora Resolution. *J. Semantics* 17, 51–89. doi:10.1093/jos/17.1.51

Esposito, A., Duncan, S., and Quek, F. (2002). "Holds as Gestural Correlated to Empty and Filled Pauses," in Proceeding of the International Conference on Spoken Language Processing (ICSLP 2002) (Denver, Colorado: ISCA), Vol. 1, 541–544.

Esposito, A., and Esposito, A. M. (2011). "On Speech and Gesture Synchrony," in *Communication and Enactment - The Processing Issues, LNCS*. Editors A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt (Stockholm, Sweden: ACL), Vol. 6800, 252–272.

Fraurud, K. (1992). *Processing Noun Phrases in Natural Discourse*. Stockholm, Sweden: Department of Linguistics - Stockholm University.

Gargiulo, C., Tronbnier, M., and Bernardini, P. (2019). The Role of Prosody in Overt Pronoun Resolution in a Null Subject Language and in a Non-Null Subject Language: A Production Study. *Glossa: A J. Gen. Linguist.*, 135–156. doi:10.5334/gjgl.973

Givón, T. (1979). *On Understanding Grammar*. New York, NY: Academic Press.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

Grönnum, N. (2009). A Danish Phonetically Annotated Spontaneous Speech Corpus (DanPASS). *Speech Commun.* 51, 594–603. doi:10.1016/j.specom.2008.11.002 Research Challenges in Speech Technology: A Special Issue in Honour of Rolf Carlson and Björn Granström

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69, 274–307.

Hirschberg, J., and Nakatani, C. (1998). "Acoustic Indicators of Topic Segmentation," in Proceedings of ICSLP-98 (ISCA: Sidney).

Kameyama, M. (1998). "Intrasentential Centering: A Case Study," in *Centering Theory in Discourse*. Editors M. Walker, A. Joshi, and E. Prince (Oxford, UK: Oxford University Press), 89–112.

Kolhatkar, V., Roussel, A., Dipper, S., and Zinsmeister, H. (2018). Survey: Anaphora with Non-Nominal Antecedents in Computational Linguistics: A Survey. *Comput. Linguist.* 44, 547–612. doi:10.1162/coli_a_00327

Krauss, R., Chen, Y., and Gottesman, R. F. (2000). "Lexical Gestures and Lexical Access: A Process Model," in *Language and Gesture*. Editor D. McNeill (Amsterdam, Netherlands: John Benjamins), 261–283.

Lyons, J. (1977). *Semantics*. Cambridge University Press, Vols. I–II.

Maclay, H., and Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *Word* 15, 19–44.

Marasovic, A., Born, L., Opitz, J., and Frank, A. (2017). "A Mention-Ranking Model for Abstract Anaphora Resolution," in Proceedings of EMNLP 2017 (Copenhagen, Denmark: Association of Computational Linguistics), 221–232.

Müller, C. (2007). "Resolving it, This and that in Unrestricted Multi-Party Dialog," in Proceedings of ACL-2007 (Prague: ACL), 816–823.

Navarretta, C. (2007). "A Contrastive Analysis of Abstract Anaphora in Danish, English and Italian," in Proceedings of DAARC 2007 Editors A. Branco, T. McEnery, R. Mitkov, and F. Silva (Lagos, Portugal: Centro de Linguistica da Universidade do Porto), 103–109.

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., and Paggio, P. (2012). "Feedback in Nordic First-Encounters: A Comparative Study,"in Proceedings of LREC 2012 (Istanbul, Turkey), 2494–2499.

Navarretta, C., and Olsen, S. (2008). "Annotating Abstract Pronominal Anaphora in the DAD Project," in Proceedings of LREC-2008 (Marrakesh, Marocco: ELRA), 2046–2052.

Navarretta, C. (2004). "Resolving Individual and Abstract Anaphora in Texts and Dialogues," in COLING-2004: Proceedings of the 20th International Conference of Computational Linguistics (Geneva, Switzerland), 233–239.

Navarretta, C. (2010). Stress, Pauses, Pronominal Types and Pronominal Functions in Danish Spoken Data. *Copenhagen Stud. Lang.*, 45–60.

Navarretta, C. (2002). The Use and Resolution of Intersentential Pronominal Anaphora in Danish Discourse. PhD thesis. Copenhagen, Denmark: University of Copenhagen.

Orăsan, C. (2003). "PALinkA: A Highly Customizable Tool for Discourse Annotation," in Proceedings of the 4th SIGdial Workshop (Sapporo, Japan: ACL), 39–43.

Paggio, P., and Navarretta, C. (2017). The Danish NOMCO Corpus of Multimodal Interaction in First Acquaintance Conversations. *Lang. Resour. Eval.* 51, 463–494. doi:10.1007/s10579-016-9371-6

Poesio, M., and Artstein, R. (2008). "Anaphoric Annotation in the ARRAU Corpus," in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) (Marrakech, Morocco: European Language Resources Association (ELRA), 1170–1174.

Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N., Roesiger, I., Roussel, A., et al. (2018). "Anaphora Resolution with the ARRAU Corpus," in Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference (New Orleans, Louisiana: Association for Computational Linguistics), 11–22. doi:10.18653/v1/W18-0702

Poesio, M. (2004). "The Mate/gnome Proposals for Anaphoric Annotation, Revisited," in Proceedings of the 5th SIGdial Workshop Editors M. Strube and C. Sidner (Cambridge, Massachusetts, United States: Association for Computational Linguistics), 154–162.

Prince, E. F. (1992). "The ZPG Letter: Subjects, Definiteness, and Information-Status," in *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*. Editors W. Mann and S. A. Thompson (Amsterdam, Netherlands: John Benjamins), 295–325.

Prince, E. F. (1981). "Toward a Taxonomy of Given-New Information," in *Radical Pragmatics*. Editor P. Cole (New York, NY: Academic Press), 223–255.

Reynolds, A., and Paivio, A. (1968). Cognitive and Emotional Determinants of Speech. *Can. J. Psychol.* 22, 164–175.

Rochester, S. R. (1973). The Significance of Pauses in Spontaneous Speech. *J. Psycholinguistic Res.* 2, 51–81.

Roesiger, I., and Riester, A. (2015). "Using Prosodic Annotations to Improve Coreference Resolution of Spoken Text," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Beijing, China: Association for Computational Linguistics), 83–88. doi:10.3115/v1/P15-2014

Roesiger, I., Stehwien, S., Riester, A., and Vu, N. T. (2017). "Improving Coreference Resolution with Automatically Predicted Prosodic Information," in Proceedings of the Workshop on Speech-Centric Natural Language Processing (Copenhagen, Denmark: Association for Computational Linguistics), 78–83. doi:10.18653/v1/W17-4610

Shriberg, E. (1994). Preliminaries to a Theory of Speech Disfluencies. PhD thesis. Berkeley: University of California.

Strube, M., and Müller, C. (2003). "A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue," in Proceedings of the ACL'03, 168–175.

Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Delogu, F., Rodriguez, K. J., et al. (2020). Annotating a Broad Range of Anaphoric Phenomena, in a Variety of Genres: The Arrau Corpus. *Nat. Lang. Eng.* 26, 95–128. doi:10.1017/S1351324919000056

Vendler, Z. (1963). The Grammar of Goodness. *Philos. Rev.*, 446–465.

Webber, B. (1988). Discourse Deixis and Discourse Processing. Tech Report. University of Pennsylvania.

Webber, B. L. (1991). Structure and Ostension in the Interpretation of Discourse Deixis. *Nat. Lang. Cogn. Process.* 6, 107–135.

Yang, B., and Mitchell, T. M. (2016). "Joint Extraction of Events and Entities within a Document Context," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (San Diego, California: Association for Computational Linguistics), 289–299. doi:10.18653/v1/N16-1033

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.