PSU McNair Scholars Online Journal

Volume 13
Issue 1 *Underrepresented Content: Original Contributions in Undergraduate Research*

Article 2

2019

Abstract Pronominal Anaphora in Three Registers of English

Dominique H. O'Donnell
Portland State University

Abstract Pronominal Anaphora in Three Registers of English

Dominique H. O'Donnell
Portland State University

2018 Ronald. E. McNair Post-Baccalaureate Achievement Program, Portland State University

Abstract

Identifying the expressions in a text that refer to the same entity, or *coreference resolution*, is an important problem in natural language processing. Abstract anaphora are distinct from other types of reference because they refer to abstract entities in discourse such as events, facts, and propositions, and their antecedents can have non-nominal phrase structure. Non-nominal antecedents are an interesting challenge in coreference resolution because the pronoun provides little information about the syntactic structure or semantics of the antecedent. A great deal of work in corpus annotation for coreference and coreference resolution has focused on newspaper text and the goal of this study is to investigate how patterns in the use of abstract pronominal anaphora vary in three text types. I compiled a corpus of newswire text, spontaneous dialog and planned speech and annotated all instances of the pronouns '*it', this',* and '*that'*. I also annotated any non-nominal antecedents used with these pronouns. I compared frequencies of these pronouns, their referential functions, and characteristics of their non-nominal antecedents. I found variation in the frequencies of referential functions, the choice of pronoun and its referential function, the grammatical structure of non-nominal antecedents and the difficulty of the annotation task. The results indicate that the range of pronominal reference, pronominal anaphora and non-nominal antecedents in spoken discourse may not be retrievable from even very large collections of newswire texts.

Introduction

Coreference resolution, identifying expressions in a text which refer to the same entity, is essential to many natural language processing tasks which rely on an understanding of discourse structure, such as text summarization and information retrieval. (Jauhar, Guerra, Gonzàlez Pellicer, & Recasens, 2015). Concrete entity anaphora are the most straightforward cases for coreference resolution and work in coreference resolution typically focuses on concrete reference (Poesio & Artstein, 2008; Jauhar et al., 2015; Kohlkatar et al., 2018). While reference to abstract entities occurs frequently in spoken registers and abstract anaphora resolution is essential for improving natural language understanding, there are few available large corpora annotated for abstract anaphora and little work in computational linguistics has considered how register effects the frequency and use of this phenomenon. The majority of available corpora annotated for coreference have been made up of primarily newspaper and newswire text. This has been the primary source of training data for coreference resolution, and it is known that newswire text differs from spoken data on a number of dimensions, including pronoun use. Lack of sufficient annotated data has contributed to a limited amount of work focused on complex types of anaphora (Poesio & Artstein, 2008).

Pronominal anaphora are pronouns which refer to an antecedent previously mentioned in discourse. *Concrete pronominal anaphora* refer to physical objects, with antecedents that surface as noun phrases. In (1), for example, the personal pronoun *it* refers to the concrete noun phrase *the car.*

(1) I parked **the car** on the hill down the street. I won't need **it** until later today.

Abstract anaphora can surface with non-nominal phrase structure and refer to abstract entities such as events, propositions or facts. In (2), for example, *it* refers to the fact-type clausal antecedent *she wasn't coming to work on Monday*.

(2) Katie told Dexter that **she isn't coming to work on Monday**. Can you believe **it**?

Abstract pronominal anaphora in English generally take the form of the personal pronoun *it* or the demonstrative pronouns *this* and *that*. There is some disagreement throughout the literature regarding the grammatical structure of abstract anaphora. Asher (1993) proposes six grammatical constructions for abstract antecedents including clauses, verb phrases and noun phrases with gerund or abstract noun heads, while Dipper and Zinsmeister (2010) suggest abstract anaphora include any anaphor with at least a verb in its antecedent. In this study, I will consider verb phrase and clausal antecedents for analysis.

The primary goal of this study is to explore how the use of abstract pronominal anaphora varies based on register. I compared the use of abstract pronominal anaphora with non-nominal antecedents in three registers of English by compiling and annotating a corpus of planned speech, conversation, and newswire texts. I found that while there are many similarities between the spoken registers, such as frequency of pronouns and anaphora, newswire text and planned speech were more alike with respect to some variables.

Literature Review

In this investigation, I am interested in abstract pronominal anaphora with antecedents that do not surface as noun phrases. I investigated the literature from the perspective of abstract anaphora and register variation in corpus linguistics and the role of corpus data for coreference resolution.

Register Variation

In this study, I will use the term register to refer to text type, based on the perspective outlined in (Biber & Conrad, 2009). Biber and Conrad explain that the analysis of texts according to register is founded on the perspective that linguistic features are functional and their frequency and distribution vary based on their function as it is influenced by discourse context and purpose. In this section, I will summarize the findings of two corpus studies that illustrate the role of register in the use of pronouns and abstract pronominal anaphora.

Byron (2003) investigated the use of all third-person pronouns in two spoken corpora, the TRAINS93 corpus task-oriented dialogs and the BUR corpus short story radio news monologues. The study analyzed pronouns and their referents in the two registers of spoken discourse and was conducted as part of a larger effort to develop automated methods for resolving reference with demonstratives and pronouns. Byron found that roughly half of the pronouns in the TRAINS93 corpus were demonstratives and half were personal pronouns. In contrast, fewer than 15% of third-person pronouns in the corpus of BUR monologues were demonstratives. Byron also found that a large number of the third-person pronouns in the TRAINS93 corpus had no linguistic antecedent because their referents were salient to both speakers. The results indicate that register plays a role in frequency and use of anaphoric pronouns.

Botley (2006) uses a corpus-based approach to investigate *indirect anaphora* in the form of demonstrative pronouns in three registers of written and spoken English. Botley defines indirect anaphora as anaphoric reference where the antecedent is not a noun phrase, the anaphor and antecedent are not coreferential, and the antecedent is not readily identifiable by the hearer or reader. Botley includes three categories of indirect anaphora: *labeling*, *situation reference*, and

discourse deixis. The study considers all demonstrative anaphora, including those with demonstrative pronouns and demonstrative noun phrases. The subcorpora represent three English registers and are comprised of samples from the Associated Press (AP) newswire text, spoken parliamentary proceedings from the Canadian House of Commons in the Hansard corpus, and literature and narrative form the American House for the Blind (APHB). Botley explains that the findings of the study show distinct patterns within each register, such as a higher frequency of retrospective labeling (anaphoric shell nouns) in argumentative genres such as metalinguistic references in parliamentary proceedings. In addition to uncovering patterns in each text type, Botley found that the task of annotating indirect anaphora is a challenge for corpus studies.

Abstract Anaphora and Coreference Resolution

In computational linguistics, anaphora resolution is generally categorized under the coreference resolution task. Much of the work in coreference resolution has focused on identifying the expressions in discourse which refer to the same concrete entity (Poesio & Artstein, 2008; Jauhar et al., 2015). In recent years, there has been increased interest in event coreference. However, this work is often limited to event coreference and event anaphora with nominal antecedents. These are the simplest case of abstract entity anaphora, where the syntactic structure is similar to that of concrete entity anaphora (Asher, 1993). Abstract anaphora with non-nominal antecedents are challenging for coreference resolution for several reasons. Their structure is complex, and it is challenging for human annotators to identify the exact boundary of a non-nominal antecedent (Kohlkatar et al., 2018). Further, there are fewer available corpora annotated for abstract reference, and lack of data has contributed to the limited amount of work focused on difficult cases in anaphora resolution, including anaphora with non-nominal or ambiguous antecedents (Poesio & Arntstein, 2008).

As resolution systems have evolved from relying on knowledge-rich, rule-based algorithms to the implementation of statistical and machine learning models, annotated corpora are commonly used as training data for these models (Poesio, Stuckardt, & Versley, 2016). While corpus studies like those of Byron (2003) and Botley (2006) indicate that the use of abstract anaphora varies across registers, there has been limited discussion of the role of register variation in selecting corpora as training and test data for resolution algorithms. Empirical analyses of the distribution of abstract anaphora can be used to inform corpus compilation and annotation for coreference resolution.

Therefore, the main purpose of this study is to investigate how the frequency of pronominal anaphora with non-nominal antecedents varies between newswire text, planned speech, and conversation.

Methods

I investigated the distribution of the pronouns *it, this,* and *that,* with nominal and non-nominal antecedents in newswire, planned speech and conversation. There are three main components to this study: corpus design and compilation, corpus annotation and analysis.

Corpus Design and Compilation

Building my own corpus was necessary to meet the goals of this study: to investigate pronoun use and characteristics of non-nominal antecedents such as structure and semantic type based on register by analyzing non-domain specific written and spoken register. While there are existing corpora annotated for coreference, they generally do not include annotations for anaphora with non-nominal antecedents. Corpora annotated for abstract reference include a limited amount of spoken data or only domain-specific speech. The annotated non-nominal antecedents in these corpora are sometimes limited to events or use markup that does not

distinguish between nominal and non-nominal antecedents. For example, commonly used corpora annotated for coreference are ARRAU and OntoNotes. Spoken data in the ARRAU corpus comes from the TRAINS-93 task-oriented dialogs and topic-specific narrative from the PEAR corpus. The English portion of OntoNotes includes 1,745,000 words across six registers, but annotations for non-nominal anaphora only include discourse deictic events (Poesio, Stuckhardt & Versley, 2016) and the annotations of non-nominal antecedents identify only their verbal head (Kohlkatar et al., 2018).

I compiled a corpus composed of three subcorpora: newswire texts from the Associated Press (apnews.com), spontaneous dialog from the Santa Barbara Corpus of spoken American English (SBCSAE), (DuBois, 2001-2005), and planned speech from a collection of TED Talk transcripts. The goals of the corpus design were to build subcorpora that are large enough to analyze a sufficient number of pronominal anaphora, and to select a variety of topics in each register so that the subcorpora were not domain specific.

I set a target of approximately 100 non-nominal anaphoric instances of *it, this,* and *that* for the final analysis. In order to estimate the appropriate size of a final corpus that could meet the target count of non-nominal anaphora, I created a small sample corpus for a preliminary investigation. The sample corpus included five 1,000 word texts for each register with newspaper text from the Wall Street Journal portion of the MASC, planned speech from TED Talk transcripts, and spoken dialog from the SBCSAE. For this preliminary investigation, I chose to annotate the middle 1,000 words of each source text. I retrieved counts for both anaphoric and cataphoric pronouns. Newspaper text included the fewest pronouns per 1,000 words with 12 total instances of *'it', 'this',* and *'that'* and 7 anaphoric instances, while TED Talk texts included 42

total instances and 30 anaphoric pronouns per 1,000 words. These findings indicated that 30,000 words of spoken discourse would include the target number of pronominal anaphora for analysis.

The final corpus is summarized in Table 1 and includes 84 texts and 79,920 tokens. I selected approximately 1,000 words from the beginning of each source text. The 27 newswire texts were retrieved from the AP news website from eight topic categories: international news, political news, US news, technology, sports, travel and lifestyle. Spontaneous dialogs in the SBCSAE were recorded in many settings across the U.S. (DuBois, 2001) and the 26 texts that I selected from the SBCSAE are limited to conversations between two or more speakers. The 31 TED Talk transcripts were retrieved from TED2SRT (ted2srt.org), a website which converts TED Talks to text files for parallel corpora. I selected TED Talks with only one speaker; multiple speaker presentations were omitted. Presentations were selected to cover a variety of topics including technology, social science, natural science, history and personal narrative.

Table 1. Corpus Composition

Register	Source	Documents	Tokens
Newswire	Associated Press	27	25,471
Planned Speech	TED Talk Transcripts	31	28,946
Conversation	SBCSAE	26	25,503
Total		**84**	**79,920**

Annotation

I developed an annotation scheme that includes five functional categories of the pronouns *it*, *this,* and *that* (see Table 2). Non-anaphoric instances were annotated as expletive *it* or as exophoric reference. Exophoric reference occurs when a referent is not linguistically introduced and can only be inferred from contextual information or shared knowledge between speakers. This includes reference to objects in a shared physical space. If a pronoun had no recoverable antecedent and was not clearly exophoric or expletive from context, it was annotated as

ambiguous. Anaphoric instances of the personal and demonstrative pronouns were coded as either concrete or abstract, and as having a nominal or non-nominal antecedent.

Table 2. Pronoun Functions

Function	Description	Example
Concrete NP Anaphor	Linguistically introduced concrete entity, NP antecedent.	a. Did you see the **red car**? b. Yes I saw **it.**
Abstract NP Anaphor	Linguistically introduced abstract entity, NP antecedent.	He told me **everything**, but I couldn't believe **it.**
Abstract Non-Nominal Anaphor	Linguistically introduced abstract entity, Non-nominal antecedent	**When the sun comes up from the horizon, the museum rises up to the sky.** **That**'s why we call it the "Aero-Solar Museum."
Exophoric Reference	Reference to entities inferable from context, no linguistically introduced antecedent.	(1) Jan talked the whole time in a voice like **this.** (2) **This** is a 3D printer. (3) **That** is a photo of me.
Expletive	Expletive *it.*	**It** is cold today.
Ambiguous Pronoun	Referent not inferable from context.	

Because nominal antecedents are not the focus of this study, nominal antecedents were not annotated. All non-nominal antecedents were annotated and assigned a unique integer value ID, and each pronoun associated with an antecedent was annotated with that antecedent's ID number. Multiple pronouns in a chain of reference share the ID number with the original mention. ID's were assigned so that membership in a chain of reference and anaphoric distance can be measured in post-processing.

Antecedents were annotated for three characteristics: structure, semantic category and ambiguity (see Table 3). *Table 3. Antecedent Annotation*

Category	Value
Structure	Sentence, multiple sentence, finite clause, non-finite clause, verb phrase, Other (prepositional phrase, adjective phrase, adverb phrase).
Semantic Category	Event, proposition, or fact.
Ambiguity	Ambiguous boundary.

The most challenging aspects of annotating non-nominal antecedents were identifying their boundaries and determining their semantic types. In the case of antecedents with ambiguous boundaries, I identified the maximum possible span of text and coded the antecedent as ambiguous. If the antecedent could be either nominal or non-nominal, and the choice was not clear from the context, the non-nominal antecedent was analyzed. There were instances where the antecedent was split and broken up by some other constituent. In these cases, the entire section of text from the first to last character was included in the annotation.

I determined the semantic type of antecedents by looking at their content and the context of both the antecedent and its referring pronouns. Due to time restraints and the complexity of identifying antecedents themselves, their semantic types were limited to *events*, *propositions* and *facts*. *Event* type antecedents were defined as any instance where the anaphor or antecedent denoted an action or event. *Fact* type antecedents included any instance where the speaker or writer reported the antecedent as factual information, while *proposition* type antecedents are questions, conditionals and antecedents or pronouns used with a stance verb.

Although there are several open-source tools available for detailed coreference annotation such as MMAX2 (Müller & Strube, 2006) and AnCoraPipe (Bertran, Borrega, Recasens, & Soriano, 2018), through the preliminary investigation I determined that the most effective approach was to use a corpus tool that I created specifically for this study. The tool traverses each element in the set of relevant word forms, ignores non-pronominal instances and selects the antecedent of each anaphor with a non-nominal antecedent for annotation. This method was efficient for the annotation task because filtering and parsing are not required in preprocessing, the candidate markables are limited to a small subset of pronouns, and layered annotations were not needed for the current study.

Analytical Procedures

The analysis focused on the use of pronouns and the characteristics of non-nominal antecedents. Annotations were added as attributes of pronoun and antecedent classes in XML and the data were parsed and output to three spreadsheets for each register. This output was used to compare frequencies in each of the three subcorpora.

The first spreadsheet listed the total count of pronouns per file and the frequency of pronoun referential functions: anaphor with non-nominal antecedent, anaphor with abstract noun phrase antecedent, anaphor with concrete noun phrase antecedent, exophoric reference, expletive and ambiguous pronouns. I calculated the frequency of these pronouns per 1000 words of each file in the corpus. I used this data to calculate the mean average frequency of pronoun functions per 1000 words in each subcorpus and to create box and whisker plots to investigate variability.

The second spreadsheet listed each annotated pronoun and its referential function. I used cross tabulation to compare the raw frequencies of individual pronouns and their referential functions in each subcorpus.

The third spreadsheet listed each antecedent's structure, its semantic type and a binary value for ambiguous boundary. I compared the raw frequencies of antecedent grammatical structures, semantic types and antecedents with ambiguous boundaries in the three subcorpora. The results of the analyses are presented in the next section.

Results and Discussion

This section reviews the frequencies of referential functions of pronouns and characteristics of antecedents. Of the grammatical features analyzed, there were no pronoun or antecedent features where mean frequencies were similar in all three registers. Similarities between the spoken registers were common for some features, while others show more similarity between planned speech and newswire text.

Referential Functions of Pronouns

The frequencies of referential functions in each register are displayed in Figure 1 and Table 6. Spoken registers have a higher frequency of anaphora overall, but conversation had the smallest proportion of non-nominal anaphora compared to other referential functions. This may be due to the high number of pronouns coded as ambiguous in conversation. The proportion of ambiguous pronouns in planned speech and newswire texts was nearly equal at 6.5% and 6.8% respectively. Planned speech had the highest proportion of anaphora with non-nominal antecedents at 26.6 % compared to newswire text at 19.7% and conversation at 18%. Although conversation included the lowest proportion of non-nominal anaphora per 1,000 words, the combined relative frequency of non-nominal anaphora and ambiguous pronouns in conversation is highest of all three registers at 36.3%. The proportion of concrete anaphora in the spoken registers was nearly equal at 42.5% and 42.2%, while concrete anaphora represented 53.8% of all pronouns in newswire text. While rare in newswire text, exophoric reference was common in

planned speech. This may be due to the amount of gestural deixis in individual talks, which featured slides and props.

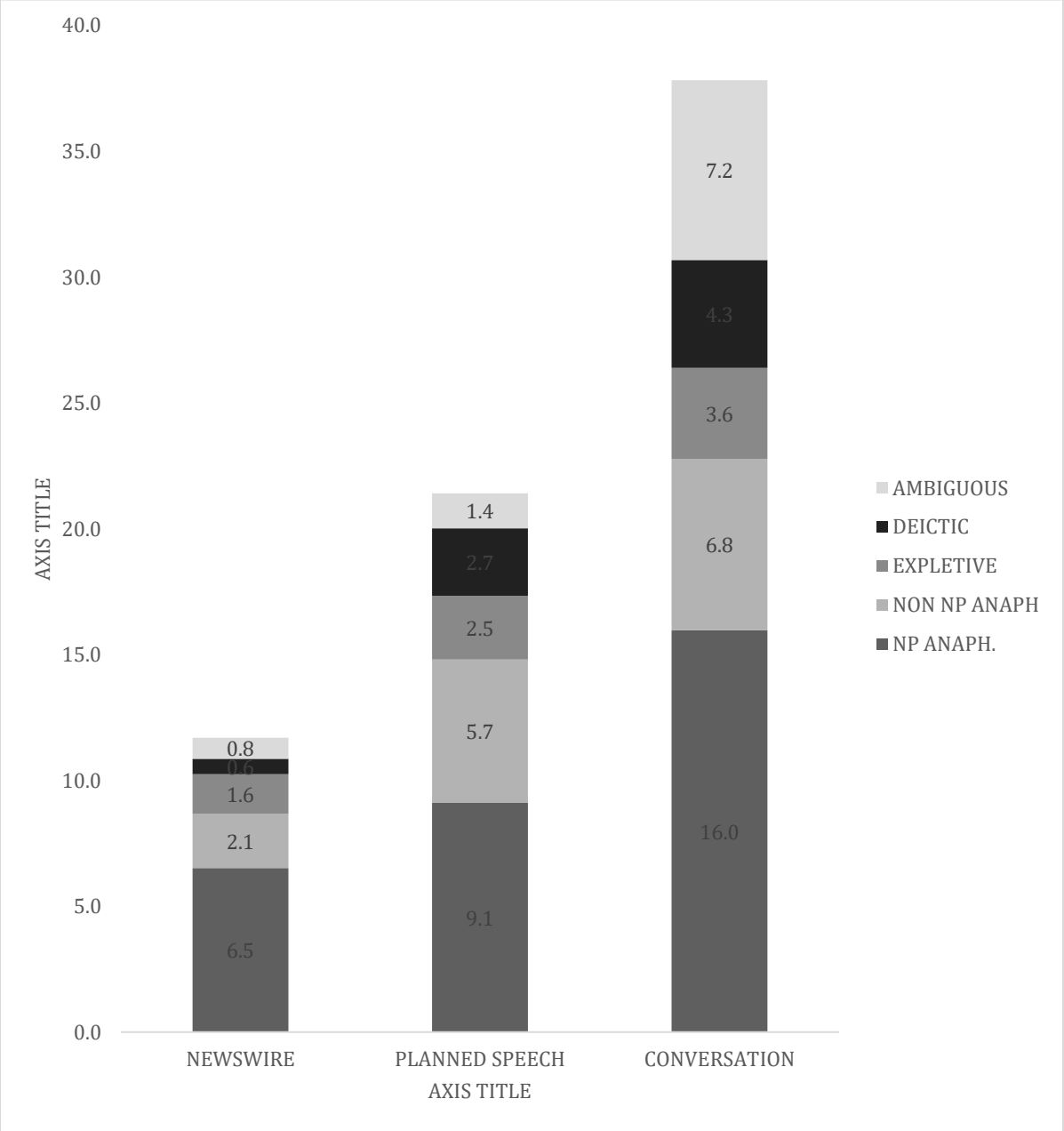

Figure 1. Mean Freq. Pronoun Usage per 1,000 Word

Table 4. Mean Freq. Pronoun Usage per 1,000 words.

Register	NP Anaphora		Non-NP Anaphora		Expletive		Exoph.Ref.		Ambiguous		Total
Newswire	6.7	**57.8%**	2.3	**19.7%**	1.6	**13.7%**	0.3	**2.6%**	0.8	**6.8%**	11.7
Planned Spch.	9.1	**42.5%**	5.7	**26.6%**	2.5	**11.8%**	2.7	**12.5%**	1.4	**6.5%**	21.4
Conv.	16.0	**42.2%**	6.8	**18.0%**	3.6	**9.6%**	4.3	**11.3%**	7.2	**18.9%**	37.8

Box and whisker plots in figures 2-6 show a different view of the frequencies of referential functions per 1,000 words in each register. In Figure 2, 'NP Anaphora' refers to the anaphora with both concrete and abstract noun phrase antecedents. Compared to the spoken registers, newswire text has the least variability for all referential functions. The median frequencies of all referential functions other than non-nominal anaphora is more similar between newswire text and planned speech than between the spoken registers. Conversation has the most variability of all three registers with respect to each of the five referential functions.

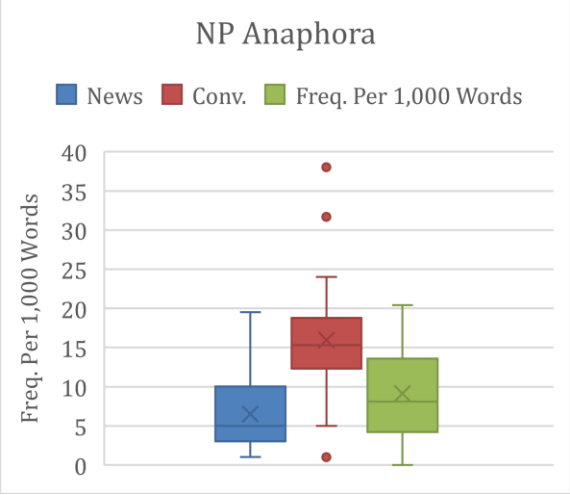

Fig. 2. NP Anaphora Freq. Per 1,000 Words

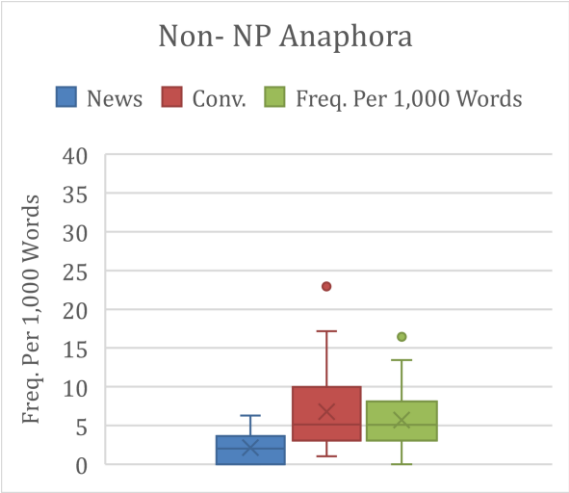

Fig. 3. Non-NP Anaphora Freq. Per 1,000 Words

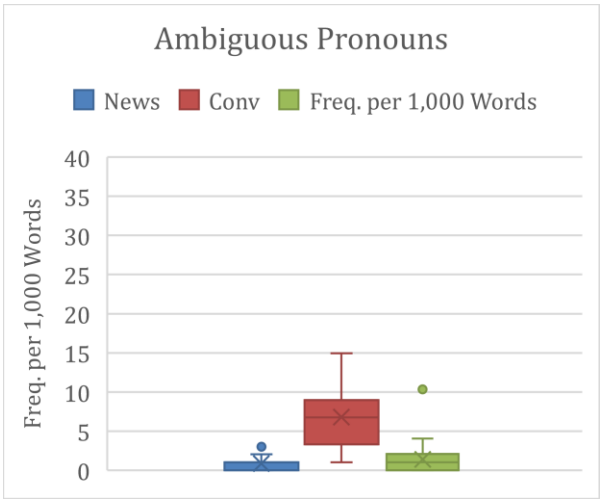

Fig. 4. Ambiguous Pronouns Freq. Per 1,000 Words

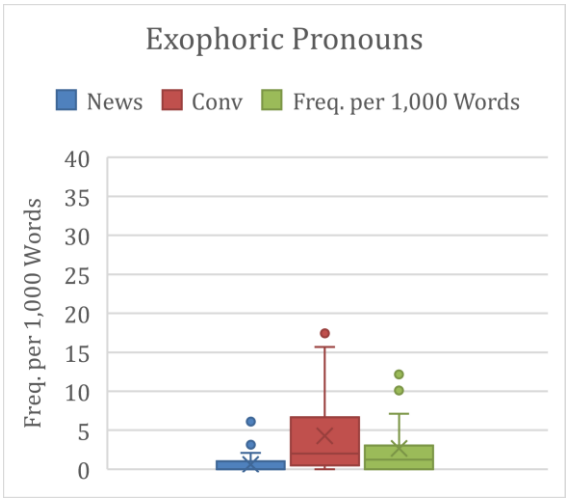

Fig. 5. Exophoric Pronouns Freq. Per 1,000 Words

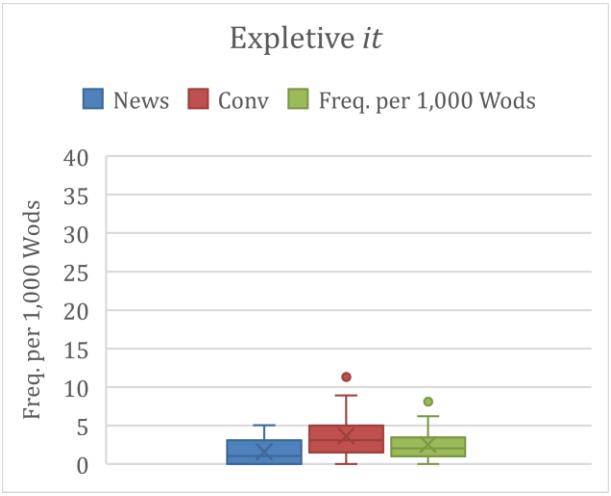

Fig 6. Expletive it Freq. Per 1,000 Words

Cross tabulation of the raw frequencies of pronouns by their referential function is summarized in Tables 5-7. Differences in the relationships between pronouns and their referential function can be seen here. The demonstrative pronoun *that* was the most common pronoun used with non-nominal antecedents in the spoken registers compared to newswire text where *that* was used nearly as often as *it. This* represented 26% of all non-nominal anaphora in planned speech compared to 15.9% in newswire text and 4.2% in conversation. In all three registers *it* was the most common pronoun overall, and it represented 82-89% occurrences of concrete anaphora.

Table 5. Cross Tabulation of Pronoun Use in Planned Speech

	Non-NP Anaphora	**Abstract NP Anaphora**	**Concr. Anaphora**	**Exoph. Ref.**	**Expletive**	**Ambig.**	**Total**
it	58	96	133	20	75	28	410 **64.9%**
this	44	6	4	55	0	6	115 **18.2%**
that	67	18	11	6	0	5	107 **16.9%**
Total	169 **26.7%**	120 **19.0%**	148 **23.4%**	81 **12.8%**	75 **11.9%**	39 **6.2%**	632 **100.0%**

Table 6. Cross Tabulation of Pronoun Use in Conversation

	Non-NP Anaphora	**Abstract NP Anaphora**	**Concr. Anaphora**	**Exoph. Ref.**	**Expletive**	**Ambig.**	**Total**
it	55	57	252	32	89	110	595 **64.1%**
this	7	1	2	42	0	6	58 **6.3%**
that	104	27	53	31	0	60	275 **29.6%**
Total	166 **17.9%**	85 **9.2%**	307 **33.1%**	105 **11.3%**	89 **9.6%**	176 **19.0%**	928 **100.0%**

Table 7. Cross Tabulation of Pronoun Use in Newswire Text

	Non-NP Anaphora	**Abstract NP Anaphora**	**Concr. Anaphora**	**Exoph. Ref.**	**Expletive**	**Ambig.**	**Total**
it	28	21	136	3	42	9	239 **79.4%**
this	10	2	5	2	0	3	22 **7.3%**
that	25	9	4	1	0	1	40 **13.3%**
Total	63 **21.0%**	32 **10.6%**	145 **48.2%**	6 **2.0%**	42 **13.9%**	13 **4.3%**	301 **100.0%**

Antecedents

Only non-nominal antecedents were annotated in this study. The relative frequency of antecedent grammatical structures varied between registers. Non-nominal antecedents with ambiguous boundaries were more common in spoken registers than in newswire text. Less variation was seen in the distribution of antecedent semantic types. *Event, proposition,* and *fact* antecedents were nearly evenly distributed in the spoken registers, but events were more common than proposition and fact in newswire text.

The raw frequency of all non-nominal antecedent structures per subcorpus is summarized in Table 8. The results show that the most common antecedent grammatical structures differ between the spoken registers. Single sentences and multiple sentences together represented 49% of antecedents in planned speech and 50% of antecedents in newswire text. These proportions are relatively high compared to 18.3% of antecedents in conversation. TED Talk transcripts used for this study were limited to talks with only one speaker and transcribed conversations selected from the SBCSAE included at least two speakers in spontaneous conversation. These texts include interrupted utterances and spontaneous shifts in topic and focus. The difference between antecedent structures in the spoken registers may be influenced by number of speakers and the careful rhetorical style of planned speech compared to spontaneous conversations.

Table 8. Raw Freq. Antecedent Structure

Register	Planned Speech		Conversation		Newswire	
Finite Clause	38	26.6%	63	48.1%	19	31.0%
Non-Finite Cl.	11	7.7%	8	6.1%	4	9.5%
Verb Phrase	18	12.6%	18	13.7%	5	7.1%
Sentence	40	28.0%	20	9.1%	18	33.3%
Multiple Sent.	30	21.0%	12	9.2%	8	16.7%
Other	6	4.2%	10	7.6%	1	2.4%
Total	143		131		55	

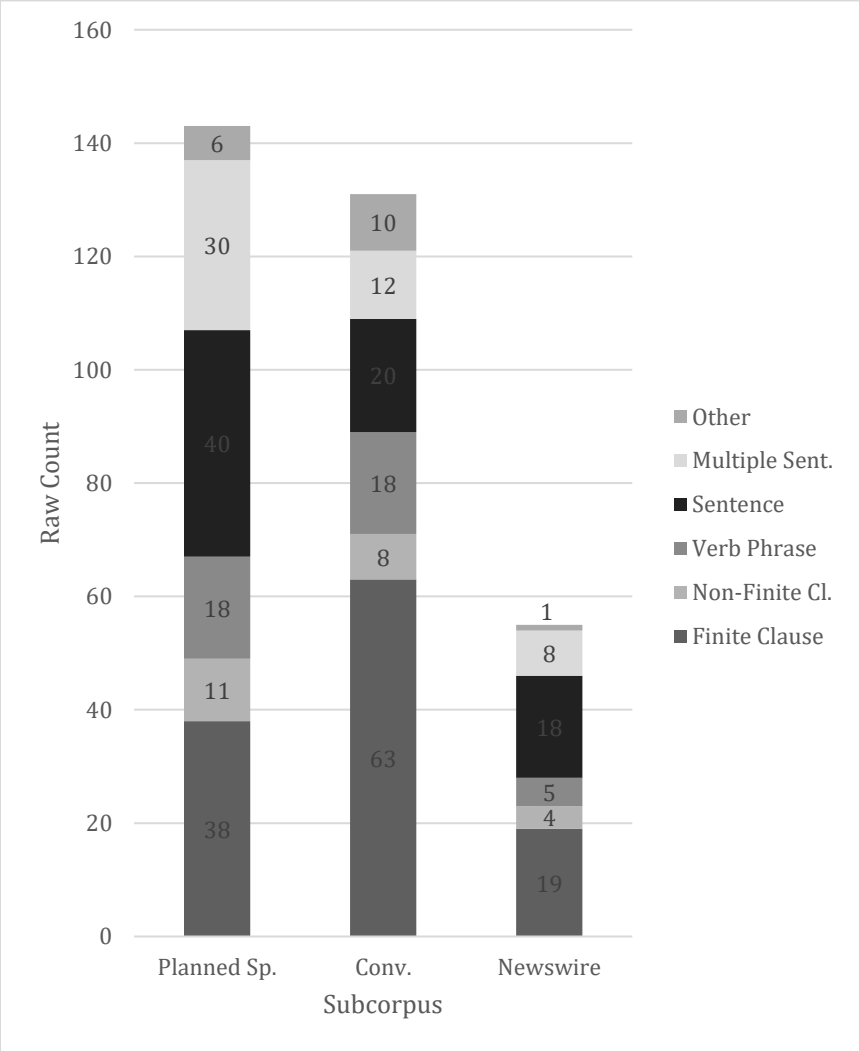

Figure 7. Raw Freq. Antecedent Structure

Table 9. Raw Freq. Antecedent Semantic Type

Register	Event		Fact		Proposition		Not Identified	
Planned Spch.	42	32.1%	44	33.6%	39	29.8%	6	4.6%
Conversation	49	34.3%	47	32.9%	45	31.5%	2	1.4%
Newswire	24	43.6%	15	27.3%	15	27.3%	1	1.8%

The raw frequencies of non-nominal antecedents with ambiguous boundaries in each subcorpus are summarized in Table 10. The results include only antecedents whose boundaries were difficult to identify, and this does not include antecedents spread across multiple turns or not bound within a single constituent. Antecedents with ambiguous boundaries were more common in spoken registers. The proportion of antecedents with ambiguous boundaries was 19.8% in planned speech and 16.1% in conversation. Only 5% of antecedents in newswire text were identified as having ambiguous boundaries.

Table 10. Non-Nominal Antecedents with Ambiguous Boundaries

Register	Non-NP Antecedents	Ambiguous Boundary	% Antecedents Ambg. Boundary
Planned Speech	143	23	16.1%
Conversation	131	26	19.8%
Newswire	42	2	5.0%

Conclusion

The aim of this study was to answer the question: How does the frequency of abstract pronominal anaphora with non-nominal antecedents vary between newswire text, planned speech, and conversation? I created three corpora of approximately 30,000 words each and annotated all instances of abstract pronominal anaphora, their antecedents, and all other uses of the pronouns *it, this,* and *that.* As expected, pronominal anaphora with non-nominal antecedents were more frequent in the spoken registers than in newswire text.

Conversation did have the highest frequency of anaphora with non-nominal antecedents and ambiguous pronouns per 1,000 words. The frequency of ambiguous pronouns in planned speech was very low and the proportion of ambiguous pronouns in planned speech was nearly identical to that of newswire text.

Cross tabulation revealed differences in the choice of pronouns used with non-nominal antecedents. *It* was the most common pronoun used anaphorically in all three registers, but an analysis of non-nominal antecedents showed that variation exists in the grammatical structure of antecedents and that the spoken registers had a significantly higher proportion of non-nominal antecedents with ambiguous boundaries.

Conversation showed more variability in the frequency of referential functions of pronouns, and further investigation is needed to determine whether there is a link between subregister and variability. The number of pronouns without identifiable referents in conversation could be affected by the number of participants and interrupted utterances. In task-oriented-dialogs, referents may be more salient to the speakers than they are in the written transcripts. The frequency of exophoric use in planned speech may have been impacted by subregister, where the use of gestural deixis increases in talks with slides, photos or props.

The analyses of pronoun function and antecedent structure showed that the frequencies of some features, such as anaphora with non-nominal antecedents and exophoric reference, were more similar between the spoken registers. The frequencies of other features, including ambiguous pronouns and antecedent grammatical structure, were between newswire and planned speech. The overall frequency of the pronouns *it, this* and *that* was higher in the spoken registers. However, compared to conversation, planned speech and newswire text included significantly fewer pronouns whose referential function was ambiguous and the relative frequency of

ambiguous pronouns. Newswire and TED Talks both include a single speaker and an expository purpose. There is also a higher degree of disfluency in spontaneous conversation compared to the careful speech used in informative presentations. However, the results do indicate that increasing the number of newswire texts in the corpus will increase the total number of pronominal anaphora with non-nominal antecedents but may not provide a sufficiently diverse set of examples representative of abstract pronominal anaphora in spoken discourse. Future research with larger subcorpora and multiple annotators is needed, but the results of this study suggest that the use of the pronouns *it*, *this* and *that* vary by register and that register variation is an important consideration in the selection of corpus data used for abstract anaphora resolution.

Limitations

This study does include a number of limitations due to time restrictions: corpus size, the annotation scheme, and no measure of inner annotator agreement. Although there were challenges in applying the annotation scheme, the existing annotations can be used to investigate the complexity of the annotation task with respect to ambiguous pronouns and non-nominal antecedents with ambiguous boundaries.

References

Asher, N. (1993). *Reference to abstract objects in discourse.* Dordrecht: Kluwer Academic.

Bertran, M., Borrega, O., Recasens, M., & Soriano, B. (2018). AnCoraPipe: a tool for multilevel annotation.

Biber, D., & Conrad, S. (2009). *Register, genre, and style.* Cambridge, UK: Cambridge University Press.

Botley, S. P. (2006). Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics, 11*(1), 73–112.

Byron, D. (2003). Annotation of pronouns and their antecedents: A comparison of two domains. Technical Report, University of Rochester.

Dipper, S., & Zinsmeister, H. (2012). Annotating abstract anaphora. *Language Resources and Evaluation, 46*(1), 37-52.

Dipper, S., & Zinsmeister, H. (2010). Towards a standard for annotating abstract anaphora. *LREC 2010 Workshop on Language Resources and Language Technology Standards*, (pp. 54–59). Valletta: ELRA.

Du Bois, J., Chafe, W., Meyer, C., Thompson, S., Englebertson, R., & Marter, N. (2000-2005). *Santa Barbara Corpus of spoken American English, Parts 1-4.* Philadelphia: Linguistic Data Consortium.

Ide, N., Baker, C., Fellbaum, C., Fillmore, C. & Passonneau, R. (2008). MASC: The Manually Annotated Sub-Corpus of American English. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech: ELRA.

Jauhar, S. K., Guerra, R., Pellicer, E. G., & Recasens, M. (2015). Resolving Discourse-Deictic Pronouns: A Two-Stage Approach to Do It. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics* (pp. 299-308).

Kolhatkar, V., Roussel, A., Dipper, S., & Zinsmeister, H. (2018). Anaphora with Non-nominal Antecedents in Computational Linguistics: A Survey. *Computational Linguistics*, 1-112. Advance online publication. doi:10.1162/coli_a_00327.

Mitkov, R. (2002). *Anaphora resolution.* London: Longman.

Müller, C., & Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun, S. Kohn, K. & Mukherjee, J. (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (p. 197-214). Frankfurt, Germany: Peter Lang

Poesio, M., & Artstein, R. (2008). Anaphoric Annotation in the ARRAU Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC),* Marrakech: ELRA.

Poesio, M., Stuckardt, R., & Versley, Y. (2016). *Anaphora Resolution*. Berlin: Springer.

Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Rodríguez, K. J., & Poesio, M. (2016). ARRAU: Linguistically-Motivated Annotation of Anaphoric Descriptions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC),* Paris: ELRA.

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., & Houston, A. (2012). OntoNotes Release 5.0. *LDC2013T19,* Philadelphia: Linguistic Data Consortium.