Variation in Coreference Patterns

Analyses across language modes and genres

Berfin Aktaş

Doctoral Thesis
submitted to the Faculty of Human Sciences,
Department of Linguistics at the University of Potsdam

in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy (PhD) in Computational Linguistics

Main Supervisor: Prof. Dr. Manfred Stede
Second Supervisor: Jun.-Prof. Dr. Tatjana Scheffler

Place and date of defense: Potsdam, April 24, 2023

Submitted on October 11, 2022
Defended on April 24, 2023

Reviewers:
Jun.-Prof. Dr. Tatjana Scheffler
Prof. Dr. Nils Reiter

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-59608
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-596086

Dedicated to my parents Güllü and Ali Haydar Aktaş

Abstract

This thesis explores the variation in coreference patterns across language modes (i.e., spoken and written) and text genres. The significance of research on variation in language use has been emphasized in a number of linguistic studies. For instance, Biber and Conrad [2009] state that “register/genre variation is a fundamental aspect of human language” and “Given the ubiquity of register/genre variation, an understanding of how linguistic features are used in patterned ways across text varieties is of central importance for both the description of particular languages and the development of cross-linguistic theories of language use.”[p.23]

We examine the variation across genres with the primary goal of contributing to the body of knowledge on the description of language use in English. On the computational side, we believe that incorporating linguistic knowledge into learning-based systems can boost the performance of automatic natural language processing systems, particularly for non-standard texts. Therefore, in addition to their descriptive value, the linguistic findings we provide in this study may prove to be helpful for improving the performance of automatic coreference resolution, which is essential for a good text understanding and beneficial for several downstream NLP applications, including machine translation and text summarization.

In particular, we study a genre of texts that is formed of conversational interactions on the well-known social media platform Twitter. Two factors motivate us: First, Twitter conversations are realized in written form but resemble spoken communication [Scheffler, 2017], and therefore they form an atypical genre for the written mode. Second, while Twitter texts are a complicated genre for automatic coreference resolution, due to their widespread use in the digital sphere, at the same time they are highly relevant for applications that seek to extract information or sentiments from users' messages. Thus, we are interested in discovering more about the linguistic and computational aspects of coreference in Twitter conversations. We first created a corpus of such conversations for this purpose and annotated it for coreference. We are interested in not only the coreference patterns but the overall discourse behavior of Twitter conversations. To address this, in addition to the coreference relations, we also annotated the coherence relations on the corpus we compiled. The corpus is available online in a newly developed form that allows for separating the tweets from their annotations.[1]

This study consists of three empirical analyses where we independently apply corpus-based, psycholinguistic and computational approaches for the investigation of variation in coreference patterns in a complementary manner. **(1)** We first make a descriptive analysis of variation across genres through a corpus-based study. We investigate the linguistic aspects of nominal coreference in Twitter conversations and we determine how this genre relates to other text genres in spoken and written modes. In addition to the variation across genres, studying the differences in spoken-written modes is also in focus of linguistic research since from Woolbert [1922]. **(2)** In order to investigate whether the language mode alone has any effect on coreference patterns, we carry out a crowdsourced experiment and

[1] `https://github.com/berfingit/TwiConv`

analyze the patterns in the same genre for both spoken and written modes. **(3)** Finally, we explore the potentials of domain adaptation of automatic coreference resolution (ACR) for the conversational Twitter data. In order to answer the question of how the genre of Twitter conversations relates to other genres in spoken and written modes with respect to coreference patterns, we employ a state-of-the-art neural ACR model [Lee et al., 2018] to examine whether ACR on Twitter conversations will benefit from mode-based separation in out-of-domain training data.

Zusammenfassung

In dieser Dissertation wird die Variation von Koreferenzmustern in verschiedenen Sprachmodi (d.h., gesprochen und geschrieben) und Textgenres untersucht. Die Relevanz der Erforschung von Variation im Sprachgebrauch wurde in einer ganzen Reihe von linguistischen Studien betont. Zum Beispiel stellen Biber and Conrad [2009] fest: "register/genre variation is a fundamental aspect of human language" und "Given the ubiquity of register/genre variation, an understanding of how linguistic features are used in patterned ways across text varieties is of central importance for both the description of particular languages and the development of cross-linguistic theories of language use."[S.23]

Wir untersuchen die Variation zwischen Genres mit dem primären Ziel, einen Beitrag zum Wissensstand zur Beschreibung des Sprachgebrauchs im Englischen zu leisten. Auf der technischen Seite glauben wir, dass das Einbeziehen von linguistischem Wissen in machine learning Ansätzen die Leistung von sprachverarbeitenden Systemen verbessern kann, insbesondere für Texte in nicht-Standard Varietäten. Neben ihrem sprachbeschreibenden Wert können die linguistischen Erkenntnisse, die wir in dieser Studie liefern, sich also als nützlich für die Verbesserung von Systemen für automatische Koreferenzauflösung erweisen; diese ist für ein tiefgreifendes Textverständnis unerlässlich, und potenziell hilfreich für verschiedene nachgelagerte NLP-Applikationen wie etwa die maschinelle Übersetzung und die Textzusammenfassung.

Insbesondere untersuchen wir ein Textgenre, das aus Konversationsinteraktionen auf der bekannten Social-Media-Plattform Twitter gebildet wird. Zwei Faktoren motivieren uns dazu: Erstens werden Twitter-Konversationen in schriftlicher Form realisiert, ähneln dabei aber der gesprochenen Kommunikation [Scheffler, 2017] und bilden daher ein für den schriftlichen Modus untypisches Genre. Zweitens sind Twitter-Texte zwar ein kompliziertes Genre für die automatische Auflösung von Koreferenzen, aufgrund ihrer weiten Verbreitung in der digitalen Sphäre sind sie aber für Applikationen, die Informationen oder Stimmungen aus den Nachrichten der Nutzer extrahieren wollen, höchst relevant. Daher sind wir daran interessiert, mehr über die linguistischen und komputationellen Aspekte der Koreferenz in Twitter-Konversationen herauszufinden. Zu diesem Zweck haben wir zunächst ein Korpus solcher Unterhaltungen erstellt und es hinsichtlich der Koreferenzbeziehungen annotiert. Wir interessieren uns dabei aber nicht nur für die Koreferenzmuster, sondern auch allgemein für diskursstrukturelle Eigenschaften von Twitter-Konversationen. Daher haben wir zusätzlich zu den Koreferenzrelationen auch die semantisch/pragmatischen Kohärenzrelationen in dem von uns erstellten Korpus annotiert. Das Korpus ist online in einer neu entwickelten Form verfügbar, die es erlaubt, die Tweets von ihren Annotationen getrennt zu repräsentieren.[2]

Diese Studie besteht aus drei empirischen Analysen, in denen wir unabhängig voneinander korpusbasierte, psycholinguistische und computerlinguistische Ansätze zur komplementären Untersuchung der Variation von Koreferenzmustern anwenden. **(1)** Zunächst führen wir eine deskriptive Analyse der Variation zwischen den Genres anhand einer korpusbasierten Studie durch. Wir untersuchen linguistische Aspekte der nominalen Ko-

[2] https://github.com/berfingit/TwiConv

referenz in Twitter-Konversationen und stellen fest, wie sich dieses Genre zu anderen Textgenres im gesprochenen und schriftlichen Modus verhält. Neben der Variation zwischen Genres steht auch die Untersuchung der Unterschiede zwischen mündlichen und schriftlichen Formen im Fokus der linguistischen Forschung beginnend mit Woolbert [1922]. **(2)** Um zu untersuchen, ob der Sprachmodus auch allein einen Einfluss auf die Koreferenzmuster ausübt, führen wir ein Crowdsourcing-Experiment durch und analysieren die Muster, die sich innerhalb desselben Genres für den gesprochenen und den geschriebenen Modus ergeben. **(3)** Schließlich untersuchen wir Möglichkeiten der Domain-Anpassung der automatischen Koreferenzauflösung für die Twitter-Konversationsdaten. Um die Frage zu beantworten, wie sich das Genre der Twitter-Konversationen zu anderen Genres im gesprochenen und geschriebenen Modus im Hinblick auf die Koreferenzmuster verhält, verwenden wir ein neuronales Koreferenzresolutionsmodell auf dem aktuellen Stand der Technik [Lee et al., 2018], um zu untersuchen, ob die Resolution auf Twitter-Konversationen von einer modusbasierten Trennung der Trainingsdaten aus externen Domänen profitiert.

Acknowledgements

The writing process of this thesis mostly coincided with the times when the COVID-19 pandemic was changing the course of life. Additionally, I embarked on a personal battle with breast cancer, which profoundly transformed me. This thesis was written in the shadow of these challenging times. Therefore, I would like to begin by expressing my gratitude to myself for the dedication and enthusiasm I displayed throughout this journey.

Being part of the A03 project team has been one of the most rewarding professional experiences I have ever had. The project meetings were always straightforward and helpful. Tatjana Scheffler has consistently provided invaluable and constructive input, always in a friendly manner. I am thankful to her for generously sharing her remarkable intellectual insights with me. The sincere conversations I had with my roommate and project companion, Yulia Clausen, will always be remembered with love and gratitude. I am indebted to Annalena Kohnert, who was a dedicated student assistant in the first years of our project, for her invaluable contributions to the annotation of our challenging corpus. Her diligent efforts and insights have been useful in enhancing the quality of our research. Being part of such an exceptional team has been a truly enjoyable experience for me.

I extend my deepest appreciation to Peter Bourgonje, Maria Skeppstedt, and Debopam Das, whose friendship has shown that true camaraderie can blossom at any stage of life. It has always been a pleasure to engage in profound discussions with this gang that span both intellectual and personal realms. I am immensely grateful for their unwavering interest in my research and their continuous support.

A special thank you goes to Michaela Schmitz and Annett Esslinger for always coming up with quick solutions to all sorts of problems. I truly appreciate their ability to find fast answers and help out whenever needed.

I would like to express my gratitude to my dear friends, Cevahir Özgüler, Buket Kaboğlu, and Elif Koca, who provided invaluable insights throughout my journey of exploration, understanding, and overcoming challenges both in the thesis writing process and in life itself. My heartfelt thanks go to the remarkable women in my life who have been a source of strength and support. Despite the physical distance, we sought solace in each other at the times of pandemic. Their persistent encouragement and genuine happiness when sharing the joys of my life will forever be etched in my memory. I extend my deepest thanks to Esma Ertürk, Esin Ertürk Asar, Nedime Ertürk, Zeynep Küçük, Sibel Hatipoğlu, Ebru Işıklı, and Dilek Akbulut Dai. I would like to express my special gratitude to Pınar Efe for wholeheartedly delving into many aspects of our experiences together. Her continuous patience and understanding have been invaluable to me in challenging times.

I also wish to express my appreciation to my flatmates, Peter Kurtenbach, Caroline Moos, and Torsten Werner, for their understanding and patience during moments when I may have been feeling uneasy and not adequately fulfilling my daily responsibilities. My sincere gratitude goes to my friends, Kıvanç Ersoy and Nabi Kımran, who have consistently stood by my side whenever I needed them. I cannot find enough words to thank my dear Sehnaz Ildan, as this thesis would not have been completed without her constant presence and unconditional support. I perceive myself as truly privileged to have her as a part of my story.

I want to express my sincere gratitude to my advisor, Manfred Stede, for his invaluable guidance throughout my academic journey. He consistently provided me with a clear understanding of what constitutes solid academic work. I always felt at ease discussing any questions or concerns with him, and together we worked collaboratively to find solutions. His critiques were always accompanied by constructive suggestions, which greatly contributed to my successful completion of the PhD process. His genuine presence has been an unwavering source of strength and has touched me deeply, both on a professional and personal level.

My dear brother Ümit Ruşen Aktaş often displayed wisdom beyond his years. Whenever I felt lost in the details, he was able to help me reconnect with reality and find my base. He meticulously read many parts of my thesis, made highly relevant suggestions, and continuously encouraged me to strive for success. I consider myself incredibly lucky to have him as my sibling, as our shared journey has allowed us to both grow and evolve together. Thank you brother!

Last but not least, I would like to express my deepest gratitude to my parents, Güllü Aktaş and Ali Haydar Aktaş, whose unconditional love and trust surrounded me. Their personal background has played a significant role in shaping my intellectual interests. The Kurdish language, which is the native tongue of both my mother and father, has never been officially acknowledged as a legitimate language within my country. Instead, it has been explicitly prohibited, and individuals who speak their native languages were subjected to severe punishment, inflicting deep wounds on the identity of our society. My contemplation of language, along with my intellectual and emotional curiosity, has arisen from witnessing the persistent determination of my people to reclaim their linguistic heritage. Over the years, I have been captivated by questions concerning the essence of language and its profound impact on our individual and collective sense of self. After a lengthy professional career in the software industry, I am happy that I have been able to commit myself to pursue my interest in working on Language. This pursuit has been made possible largely due to the unwavering support and encouragement of my family.

This thesis is dedicated to my parents for their unconditional love and to those who endured childhoods oppressed by language bans yet demonstrated incredible resilience.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Referring in Speaking and in Writing

(1.1) Paul tried to call George on the phone, but he wasn't successful.

When an English-speaking person hears or reads this sentence[1] (1.1), (s)he will probably **understand** what it means. The sentence contains a number of noun phrases: *Paul*, *George*, *the phone*, and *he*. Inferring what *he* refers to, also known as the process of **pronoun resolution**, is an essential step in understanding the sentence and humans do it automatically. Although there are two candidates in this sentence, the human reader can immediately disambiguate the potential expressions and infer that *Paul*, not *George*, is what *he* refers to. As a result, *he* and *Paul* are **referring expressions** that refer to the same real world entity, so they **corefer**.

Disambiguation of potential coreferring expressions in 1.1 can not be achieved by only using the morphosyntactic information as both *George* and *Paul* agree in number and gender. One can hypothesize that a simple grammatical heuristic might work (i.e., *Paul* and *he* are both subjects of the clauses they belong to, so it is more likely that *he* refers to *Paul.*). But let us look at another sentence (1.2) that shares the same grammatical structure as the first one and differs from it only by one word:

(1.2) Paul tried to call George on the phone, but he wasn't available.

In this case, *he* refers to *George* instead of *Paul.* The heuristic of grammatical parallelism is ineffective for this instance. This sentence pair represents a hard case of pronoun resolution, where referential patterns cannot be adequately explained by morphosyntactic or structural heuristics. It is necessary to use background knowledge not expressed in the sentences' words to resolve the pronoun. This and numerous other cases indicate that resolution of coreferring expressions (i.e., **coreference resolution**) is not a trivial task in natural language processing.

Referring expressions have a wide range of forms, including pronouns, names, and definite or indefinite noun phrases. *Referential choice*, the process of choosing an appropriate expression for referring to the entity of concern, is influenced by the cognitive status of the entity in the text, that is characterized by, for instance, its givenness, topicality, accessibility, salience, and prominence (e.g., Chafe [1976], Givón [1983], Ariel [1990], Gundel et al. [1993], von Heusinger and Schumacher [2019]). In most of these theoretical approaches, textual distance between the referring expressions is described as a key component for the evaluation of the cognitive status of the entity.

[1] Examples 1.1 and 1.2 are taken from a Winograd Schema Challenge dataset: `https://huggingface.co/datasets/winograd_wsc`

Research on coreference indicates that the genre and language mode (spoken vs written) of a text have influence on referential choice and, in turn, referential patterns. For instance, Fox [1987] analyzes the use of the third person singular pronouns in spoken and written texts and discovers that spoken texts allow for a greater textual distance between pronouns and an earlier expression that refers to the same entity (i.e., anaphoric distance). In another line of research in linguistics, when learning-based automatic coreference resolution (ACR) systems are tested on text varieties that are different from the training data in terms of genre or mode, their performance drops dramatically. Therefore, such computational experiments also confirm the different distributions in coreference features across different text varieties.

In the 1970s and early 1980s, empirical corpus-based research examining the spoken-written discrepancy discovered that texts created in different modes could occasionally be more *similar* in terms of relevant linguistic features than texts generated in the same mode. Based on this empirical observation, a number of influential linguists contend that rather than being considered as mutually exclusive concepts, spoken and written language should be viewed as a **continuum** that allows various text genres to have varying degrees of "spokenness" and "writtenness" Tannen [1982a], Koch and Oesterreicher [1985], Biber [1988]. Researchers acknowledging the continuum approach sometimes have theoretical engagements that associate certain linguistic features with certain situational characteristics, such as immediacy and distance in language by Koch and Oesterreicher [1985].

Fox [1987] argues that in order to examine the impact of the language mode in a comparative study, typical genres for both modes should be included in the analysis:

> there are marked and unmarked types of text for each modality, and that we should compare texts of the same markedness (see Biber 1983). For example, multi-party spontaneous conversation is the unmarked text-type for oral production, and expository prose is one of the unmarked text-types for written (what Biber calls 'literate') production. This way of looking at genre and modality seems to me to be highly appropriate, since it allows us to compare what one typically does when writing with what one typically does when speaking. [p.138]

In contrast to Fox [1987], Tannen [1982b] and Akinnaso [1982] argue that contrastive studies that use different genres for comparing the spoken-written modes can occasionally be misleading. Tannen [1982b] states that comparing different genres from written and spoken modes may not provide information about the impact of the mode on language because "[..] it is impossible to determine whether such findings reflect the spoken vs written modes or other aspects of the data-such as genre, or context and associated register" [Tannen, 1982b, p.6]. Akinnaso [1982] makes a suggestion about what constitutes comparable texts across the written-spoken distinction. Akinnaso [1982] states that to be comparable, the two texts must be produced by the same person, same degree of planning, same degree of formality, same degree of interactiveness (i.e., both must be monologue or both must be dialogue), and be of the same genre (story or essay, for example).

While acknowledging the continuum approach, we also believe that both modes could have characteristics that are unique to them. Consequently, analyzing mode differences is equally important. We recognize the opposing viewpoints on the methodology of research on mode differences, such as those expressed by Akinnaso [1982] and Fox [1987]. In our research, we consider these approaches in a complementary manner, which is briefly described in the following section.

1.2 Motivation

This thesis explores variation in coreference patterns across language modes (i.e., spoken and written) and text genres. The significance of research on variation in language use has been emphasized in a number of linguistic studies. For instance, [Biber and Conrad, 2009, p.23] state that "register/genre variation is a fundamental aspect of human language" and "Given the ubiquity of register/genre variation, an understanding of how linguistic features are used in patterned ways across text varieties is of central importance for both the description of particular languages and the development of cross-linguistic theories of language uses".

We examine the variation across genres with the primary goal of contributing to the body of knowledge on the description of language use in English. On the computational side, we believe that incorporating linguistic knowledge into learning-based systems can boost the performance of automatic natural language processing systems, particularly for non-standard texts. For this type of further investigation, in addition to its descriptive value, the linguistic findings we provide in this study may prove to be helpful for the improvement of the performance of automatic coreference resolution systems.

This study consists of three empirical analyses in which we independently apply corpus-based, psycholinguistic and computational approaches for investigating our research interests. We believe these approaches complement one another in our analysis.

Corpus-based analysis: In response to (i) inconclusive results in the literature as to the properties of coreference in written versus spoken language, and (ii) a general lack of literature on automatic coreference resolution on both spoken language and social media, we undertake a corpus study involving various genres. Digital texts are one sort of text variety that has been claimed to combine elements of spoken and written language (e.g., Jonsson [2016]). In our case, in addition to several spoken and written genres, we are interested in the characteristics of conversational texts from Twitter[2], a very popular social media platform that enables near-synchronous communication[Scheffler, 2017] among its users. We compare the texts in terms of distance-based (i.e., the linear textual distance between the referring expressions), frequency-based (i.e., the frequency distribution of referring expressions in terms of their grammatical categories) and heaviness-based (i.e., heaviness of the linguistic forms of referring expressions) characteristics. We adopt a data-driven, theory-neutral approach, and do not aim to interpret our results in terms of theoretical engagements on genre variations, such as language of immediacy or distance proposed by [Koch and Oesterreicher, 2012]. We believe that our analyses can be used to significantly improve the automatic coreference resolution, especially when it is applied to more diverse data that is common today. In addition, results of our comparative corpus study is relevant for the automatic identification of text genres in terms of coreference patterns.

The genres in our corpus-based study include both those that are typical (i.e., unmarked in Fox [1987]'s terminology) of their respective modes (e.g., telephone conversations for spoken communication, news for written communication) as well as some that are atypical. For example, broadcast news, which was initially created in written form but is performed orally, and Twitter texts, which are realized in written form but resemble spoken communication [Scheffler, 2017], are included in the analysis. As a result, we have the opportunity to examine how typical and atypical genres of their modes are positioned in spoken-written continuum.

[2] https://twitter.com/

Story continuation experiment: Even though the idea of continuum is useful when exploring and analyzing spoken and written texts, it is still intriguing to look at how different modes affect coreference patterns. Following Akinnaso [1982]'s proposal about comparable texts, we carry out an experimental study to investigate the coreference features in spoken and written modes for the same genre. As noted earlier, corpus studies demonstrate that spoken genres include longer anaphoric distances than written genres, if the distance is measured in terms of number of clauses (e.g., Fox [1987]). We design a story continuation experiment where we systematically manipulate the anaphoric distance in order to examine the differences in spoken and written modes. We aim to gain more insight about the impact of the mode, in a setting where the textual differences (e.g., clause length and clause types) are eliminated, when both modes convey a similar level of spontaneity, informality and interactivity. To our knowledge, this is the first study of its kind, where anaphoric distance is manipulated systematically in a language production experiment in order to isolate and examine the effect of modes.

Computational experiments: Performances of automatic coreference resolution methods are known to drop significantly on non-standard texts, such as spoken or online conversations (e.g., Khosla et al. [2021], Dakle et al. [2020]). Tweets are especially challenging for ACR because they often contain, for instance, non-standard words such as the ones beginning with # and @, exophoric pointers to non-linguistic content and mixed pronominal references to the same entity due to the nature of multi-user conversations. The resolution of coreference links in tweets, however, is highly relevant for many applications that seek to extract information or sentiments from users' messages. We explore the domain adaptation of ACR for the conversational Twitter data in order to enhance its performance on this genre. Since we are interested in the question of how the medium of *microblog*, particularly Twitter, relates to the spoken-written continuum with respect to coreference patterns, we experimentally examine, with a state-of-the-art neural ACR model, whether ACR on Twitter conversations will benefit from mode-based separation in out-of-domain training data. To the best of our knowledge, the genre of Twitter conversations has not yet been examined in the scope of automatic coreference resolution research.

1.3 Contributions and Summary

This thesis has two main objectives:

1. We explore how the language in spoken and written modes, and the text genres generated in these modes, relate to each other in terms of coreference patterns.
2. We examine the genre of Twitter conversations in terms of linguistic and computational aspects of coreference and how this genre relates to other text genres in spoken and written modes. This genre remains understudied by the coreference research literature.

For addressing these objectives:

- We compile a new corpus composed of conversational Twitter data (i.e., the TwiConv corpus) and manually annotate it for coreference. According to some researchers (e.g., Fox [1987]), coreference patterns are associated with discourse-structural characteristics. To facilitate further study into this association, we additionally annotate discourse relations in the TwiConv corpus.[3]

[3] The corpus is available online in a newly developed form that allows for separating the tweets from their annotations: `https://github.com/berfingit/TwiConv`

- We examine how the genre of Twitter conversations relates to other genres in the well-known OntoNotes and Switchboard corpora in terms of linguistically motivated coreference features. With this comparative corpus study, we explore how the examined genres, especially Twitter texts, are placed in the spoken-written continuum in terms of characteristics of coreference patterns.

- We improve the performance of a state-of-the-art coreference resolution system [Lee et al., 2018] by 21.6% F1 score, which is originally trained on OntoNotes, by retraining on our manually-annotated TwiConv data. Further experiments by combining different portions of OntoNotes with TwiConv in training data reveal how the language used in Twitter conversations relates to spoken-written language in terms of computational aspects of coreference.

- We conduct a story continuation experiment to examine the impact of language mode on coreference patterns for the same genre. The experiment is carried out via crowdsourcing.

1.4 Thesis Outline and Corresponding Publications

To set the stage for the remainder of the thesis, **Chapter 2** provides some theoretical background on linguistic and computational aspects of coreference, clarifies the terminology, and reviews prior work in the field.

Chapter 3 introduces TwiConv, an English coreference-annotated corpus of microblog conversations from Twitter. We describe the corpus compilation process, the annotation procedure, and the corpus distribution method. In addition to TwiConv, we also provide details on the widely used coreference-annotated corpora (OntoNotes and Switchboard) that we use in our empirical work.

Corresponding publication:

- **Berfin Aktaş** and Annalena Kohnert. 2020. TwiConv: A Coreference-annotated Corpus of Twitter Conversations. In Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference, pages 47–54, Barcelona, Spain (online). Association for Computational Linguistics.[4]

Chapter 4 describes the comparative corpus-based study we carry out. We examine the texts through descriptive and explorative (i.e., hierarchical cluster analysis) statistical methods and report our findings referring to the continuum approach.

Corresponding publication:

- **Berfin Aktaş**, Tatjana Scheffler, and Manfred Stede (2019). Coreference in English OntoNotes: Properties and Genre Differences. In: Ekštein, K. (eds) Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science, vol 11697. Springer, Cham.

- **Berfin Aktaş** and Manfred Stede. 2020. Variation in Coreference Strategies across Genres and Production Media. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5774–5785, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chapter 5 describes the story continuation experiment we design in order to examine the differences in spoken and written modes, using the same genre. The chapter presents the

[4] Annalena Kohnert implemented the corpus distribution method, hence she authored most of the relevant section in the paper.

participant qualification procedure as well as the description of the experiment and its findings obtained through descriptive and inferential statistics (i.e., Generalized Mixed-Effects Logistic Regression). We also discuss the results of preliminary classification experiments we employed on this data.

Corresponding publication:

- **Berfin Aktaş** and Manfred Stede. Anaphoric distance in oral and written language: Experimental evidence. Discours, Issue 31, 2022. (Accepted)

Chapter 6 presents the computational experiments conducted via a state-of-the-art ACR system [Lee et al., 2018]. The experiments aim to improve the performance of ACR on conversational Twitter data, as well as to reveal how Twitter conversations relate to other spoken and written genres, in terms of the computational features of coreference.

Corresponding publication:

- **Berfin Aktaş**, Tatjana Scheffler, and Manfred Stede. 2018. Anaphora Resolution for Twitter Conversations: An Exploratory Study. In Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference, pages 1–10, New Orleans, Louisiana. Association for Computational Linguistics.

- **Berfin Aktaş**, Veronika Solopova, Annalena Kohnert, and Manfred Stede. 2020. Adapting Coreference Resolution to Twitter Conversations. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2454–2460, Online. Association for Computational Linguistics.[5]

Chapter 7 provides context for the relevancy of investigating coherence and coreference together, and describes the annotation procedure for coherence relations on the TwiConv corpus[6].

Corresponding publication:

- Tatjana Scheffler, **Berfin Aktaş**, Debopam Das, and Manfred Stede. 2019. Annotating Shallow Discourse Relations in Twitter Conversations. In Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, pages 50–55, Minneapolis, MN. Association for Computational Linguistics.[7]

Chapter 8 summarizes our findings and explores further directions.

[5] This is a joint work with the graduate student Veronika Solopova and our student assistant Annalena Kohnert. I conceptualized the study, including the design of the computational experiment settings, the specifications for the test-train datasets, and the interpretation of findings. The experimental setups and the experiments were carried out by Veronika Solopova. The scripts produced by Annalena Kohnert are used for the analysis in the paper's final part.

[6] Annotation statistics are computed using the scripts implemented by another graduate student, Burak Özmen.

[7] This article is related to our coherence relation annotations on the TwiConv corpus. I coordinated the annotation process, including the inter-annotator agreement study and contributed the paper by providing annotation and inter-annotator agreement statistics, and writing the relevant section.

2

Background

In this chapter, we give a theoretical background for the remainder of the work, as well as the review of relevant research from the literature. We first present information on the linguistic and computational aspects of coreference in Section 2.1, including details on common coreference-annotated corpora and automatic coreference resolution approaches. Then, in Section 2.2, we briefly describe what Twitter is, as well as how it functions and how the users interact with it. In Section 2.3, we clarify the terminology we use concerning the linguistic variation and provide a brief summary of the research on the distinctions between spoken and written modes. Finally, we provide a comprehensive summary of the comparative research on coreference patterns.

2.1 Coreference

When humans communicate —whether orally, in writing, or through sign language— they do so through a coherent whole rather than through independent clauses or sentences. Coherence and cohesion are the linguistic phenomena that account for this assumption. Coherence is defined in terms of the underlying semantic relationships between the units of text, whereas cohesion holds between the surface elements of the text. [Stede, 2011, p.20] defines cohesion as a phenomenon of 'textuality' indicating that "sentences and clauses are connected to one another by a variety of linguistic means that can be identified at the text surface". **Anaphora** is a type of cohesion that points back to a previous item in the text [Halliday and Hasan, 1976]. The linguistic element (e.g., word or phrase) that points back is called the **anaphor** and the element it points to is the **antecedent**. An anaphoric relation between two elements indicates that the anaphor can be fully interpreted only in the context of an antecedent. Identification of the antecedent for an anaphor is called **anaphora resolution**.

Typical anaphors in language are pronouns for which, in most cases[1], the antecedent should be known in order to get the meaning of the pronoun as illustrated in 2.1[2]. In this example, the anaphor ("she") and its antecedent ("the cyclist who won the race") **denote** the same real-world entity. Therefore, the relation between these two elements is called as an **identity relation**.

Frege [1892] makes a distinction between two dimensions of linguistic meaning, "Bedeutung" and "Sinn", which Max Black later referred in English as "reference" and "sense", respectively. The **reference** of a nominal expression, in this distinction, is the relation

[1] Exceptional cases are the deictic pronouns which we mention below.

[2] In the example, both the pronoun "she" and its antecedent "the cyclist who won the race" are put in the square brackets and marked by the same subscript "i". This notation indicates that these two items are denoting the same entity. We will use the same notation throughout the thesis to mark the similar identical instances.

with the entity the expression denotes (i.e., related to the extra-linguistic context) while the sense describes the way it presents the reference. In accordance with this terminology, if an expression denotes a discourse entity[3], it is a **referring** (or **referential**) **expression**. Two elements in a discourse denoting the same entity have identical **referent** and they are termed **coreferential**. The linguistic phenomenon **coreference** accounts for the act of denoting the same referent. Referring expressions pointing to a discourse referent are called the **mentions** of that referent. The set of all mentions of the same entity forms a **coreference chain** or **coreferential chain**. In 2.2[4], the following coreferential chains that are also differentiated by the subscripts in the notation, are distinguishable: {Sophia Loren, she, the actress, her, she} and {Bono, the singer}. In addition to the mentions in these chains, "a thunderstorm" and "a plane" are also mentions of real world entities but the entities denoted by these two expressions are referred to only once in the text. The sets containing these mentions are termed **singleton**, which can be described as a coreference chain with only one element. Identification of the coreference chains in a text is called **coreference resolution**.

(2.1) I met [the cyclist who won the race]$_\mathrm{i}$. [She]$_\mathrm{i}$ deserved that result

(2.2) [Sophia Loren]$_\mathrm{i}$ says [she]$_\mathrm{i}$ will always be grateful to [Bono]$_\mathrm{j}$. [The actress]$_\mathrm{i}$ revealed that [the U2 singer]$_\mathrm{j}$ helped [her]$_\mathrm{i}$ calm down when [she]$_\mathrm{i}$ became scared by [a thunderstorm]$_\mathrm{k}$ while travelling on [the plane]$_\mathrm{n}$.

Anaphora and coreference, as indicated above, are closely related concepts but they are not the same. For instance, an anaphora relation can be established between two elements which are not referring to an identical entity, and hence, are not coreferential. For instance, in 2.3[5], the phrase "The engine" is an anaphor because that can only be fully interpreted in the context of "a black Mercedes". However, its relation to the antecedent "a black Mercedes" is not an instance of identity relation. The engine is the part of the real world entity "a black Mercedes"; therefore, its relation with the antecedent is termed a "part-of relation". Part-of relation is an instance of non-identity anaphoric relations that are often called **bridging relations** in the literature. In this thesis, we are interested only in identity relations. Therefore, the bridging relations are out of our scope. For discussions on the theoretical and practical aspects of bridging relations, we refer to Clark [1977], Wei [2014], Roesiger [2018], Roesiger et al. [2018], Hou et al. [2018].

(2.3) I saw *a black Mercedes* parked outside the restaurant. The engine was still running.

Not all coreferential relations have to also be anaphoric, just as not all anaphoric relations have to be coreferential. For instance, in 2.4 the two instances of "the Pope" refer to the same real-world entity and therefore, they are coreferential. However, even in the absence of the first instance of "the Pope", the second instance can still be clearly interpreted. As a result, they are coreferential but not anaphoric.

(2.4) On Monday, [the Pope]$_\mathrm{i}$ went to Brazil, and on Wednesday, the plane with [the Pope]$_\mathrm{i}$ landed in Argentina.

The referring expressions we illustrated in the examples so far have different forms (e.g., pronoun, definite noun phrase, proper name). But they are all nominal expressions. Coreference occurring only among the nominal expressions is called **nominal coreference**. There can also exist non-nominal referring expressions. For instance, in 2.5[6] the

[3] That can be a real-world entity or an abstract object present in the processed discourse.

[4] The example is taken from [Mitkov, 2002, p.5].

[5] Taken from [Poesio et al., 2018, p.13]

[6] Taken from [Poesio et al., 2018, p.14]

demonstrative pronoun “that” refers to the syntactically non-nominal expression “we ship one boxcar of oranges to Elmira”. In this study, we are only interested in the nominal coreference. Therefore, relations established by non-nominal referring expressions are out of the scope of our work (see Kolhatkar et al. [2018] for a survey on non-nominal anaphora).

(2.5) so we ship one boxcar of oranges to Elmira and <u>that</u> takes another 2 hours

Pronominal anaphora is the most studied form of anaphora. Pronouns, especially third person pronouns, are mostly used as anaphors such as “she” and “her” in 2.2. However, there also exist the reference type of **exophora**, in which pronouns are used to refer to extra-linguistic context. For instance, **deictic usages** (i.e., **deixis**) are instances of exophora, in which no explicit antecedent is present in the text; instead the referent is conceived from the context of the utterance or speaker. The first and second person pronouns are common examples of deixis as their interpretation depends on, for instance, the extra-linguistic information of who is uttering the expression. There exist other non-anaphoric instances of pronouns, also referred to as **pleonastic** usages. The third person pronoun “it” appearing, for instance, in the descriptions of weather conditions (e.g., *It* is sunny.) or in idiomatic contructions (e.g., We made *it*!) are instances of pleonastic use.

Although coreference and anaphora can be distinct, they are often used interchangeably by linguists. This is partly because by far the most studied type of anaphoric reference is coreferential (e.g., pronominal anaphora). In this thesis, we use the terms interchangeably only when we are referring to the intersection of these two linguistic phenomena. Otherwise, coreference is what interests us here.

Since we are doing an empirical study of coreference in this thesis, we present relevant material in the rest of this chapter. First, we give a brief overview of the existing corpora that are manually annotated for coreference and then we make an overview of research on the automatic coreference resolution task.

2.1.1 Coreference-annotated Corpora

For the empirical examination of coreference, there are manually annotated corpora available. Among these, MUC [Hirschman and Chinchor, 1998] is regarded as the first comparatively sizable coreference-annotated corpus. Since then, several additional corpora have been made available, with diverging labeled constituent types (i.e., **markables**) and referential relation types (e.g., identity relations, bridging anaphora). Several coreference-annotated corpora that were extensively used in linguistic and computational investigation of coreference phenomena are briefly described here.

MUC : MUC is used to name a group of corpora that have been originated in response to the needs for the tasks specified in the **M**essage **U**nderstanding **C**onferences (MUC-6 and MUC-7) [Grishman and Sundheim, 1996, Hirschman and Chinchor, 1998]. MUC are the first corpora that are used for the large-scale evaluation of coreference systems. The MUC-6 and MUC-7 datasets are based on newspaper texts in English from the Wall Street Journal and the New York Times, respectively. Only the identity relations are annotated in the MUC corpora. Markables in the MUC annotations are nominal expressions. Singletons are not annotated in the MUC corpora. MUC-6 and MUC-7 contain 3K markables each organized in 60 and 50 documents, respectively[7].

ACE ACE corpus has been originated with the **A**utomatic **C**ontent **E**xtraction conferences [Doddington et al., 2004]. ACE data includes broadcast news and newspaper

[7] The numbers are computed from the statistics in [Hoste, 2005, p.21]

articles in English, Chinese and Arabic. The identity relations are annotated between the nominal mentions. Different from MUC, only the mentions referring to the certain entity types (person, organization, geo-political entity, location, facility, vehicle, and weapon) are considered in ACE and singletons are also marked. In addition to the identity relations, there also exist annotations for the cases of metonymy, which happens when the name of one entity is used to refer to another entity (or entities) that is closely associated with it, for instance, when "the White House" is used to refer to the "Administration of the United States". The dataset is composed of 350K tokens in total[8].

Switchboard Switchboard is a long standing corpus of conversational speech with distinct annotation layers for syntax, speech act, information status, coreference and other linguistic concepts [Godfrey et al., 1992]. Calhoun et al. [2010] brought together the existing annotations on Switchboard corpus and delivered a combined resource that contains 642 dialogs. 147 of these dialogs have annotations for coreference. Coreference annotation was carried out as part of the information status annotation in Switchboard. As information status is a property of entities, only nominal expressions were considered as markables and only the anaphor-antecedent pairs of the entities with a certain information status value are labeled. Similar to MUC, singleton mentions are not annotated in Switchboard. The corpus is composed of 240K tokens. For a more detailed review on Switchboard corpus, see Section 3.2.2.

OntoNotes OntoNotes is comprised of various varieties of text such as telephone conversations, broadcast conversations, and news in English, Arabic and Chinese Weischedel et al. [2013]. Unlike ACE or Switchboard, coreference annotations are not restricted to the specific kind of entities; instead all types of nominal expressions are annotated, regardless of the entity type or information status. Nominal expressions are considered as markables in OntoNotes coreference scheme [BBN Technologies, 2007]. In addition, verbs are also considered as single-word mentions if they corefer with a noun phrase as in the example 2.6 that is taken from [Weischedel et al., 2013, p.20]. Similar to MUC and Switchboard, singleton entities (i.e., entities referred only once) are not annotated in OntoNotes.

(2.6) Sales of passenger cars [grew]$_i$ 22%. [The strong growth]$_i$ followed year-to-year increases.

In addition to coreference, OntoNotes also contains various other gold annotation layers such as PoS tagging, constituency parsing and word sense annotation. A portion of OntoNotes which contain all of these annotation layers, with 1.6M English words, 950K Chinese words, and 300K Arabic words,is used as a reference dataset for evaluating the competing systems in the CoNLL-12 shared task Pradhan et al. [2012]. That CoNLL shared task data has in recent years become a standard data benchmark for evaluating and comparing the performance of the automatic coreference resolution systems. For a more detailed review of OntoNotes, see Section 3.2.1.

ARRAU Similar to OntoNotes, ARRAU [Uryupina et al., 2016, Poesio et al., 2018] contains texts from different text varieties such as task-oriented dialogues, narratives from English Pear Stories corpus [Chafe, 1980], and newspaper texts from the Wall Street Journal. All noun phrases, including the singletons, are treated as markables. In addition to the identity relations, discourse deixis (e.g., example 2.5) and bridging relations (e.g., example 2.3) are also annotated in ARRAU. The corpus is composed of 340K tokens

[8] Statistics regarding the annotations were not provided in the reference paper of Doddington et al. [2004]

arranged in 552 documents. Annotations contain 99K markables in total, in which 1.6K are discourse deixis and 5.5K are bridging anaphora instances[9].

2.1.2 Coreference Resolution

Coreference Resolution (CR) is the identification of all expressions referring to the same referent. Resolution of coreferential elements is essential for a good text understanding. It has also been demonstrated to be beneficial for several downstream NLP applications, including machine translation [Lapshinova-Koltunski et al., 2018] and text summarization [Steinberger et al., 2007].The linguistic and NLP communities have therefore been focusing on this task for a long while.

The task of automatic coreference resolution (ACR) is a challenging problem and, although it has a long history (see the details in 2.1.2.1), it is still far from being fully resolved. We illustrate a complicated case of CR in example 2.7, taken from [Ng, 2017, p.4877]. In the example, the possible antecedents for the first occurrence of **her** are **The Queen Mother** and **Queen Elizabeth II**, but the text does not provide sufficient information for distinguishing the correct referent. In order to properly resolve the pronoun, extralinguistic information that Princess Margaret is Queen Elizabeth II's sister should be provided. For the second occurrence of **her**, again extralinguistic common sense knowledge that "it does not make sense for Person A to summon Person B to treat Person B's problem" [Ng, 2017, p.4877] is employed to rule out **Nancy Logue** as an antecedent. While most human resolvers can identify the correct antecedents without any problem, these cases are challenging for automatic CR systems.

(2.7) The Queen Mother asked [Queen Elizabeth II]$_i$ to transform [her]$_i$ sister, [Princess Margaret]$_j$, into a viable princess by summoning a renowned speech therapist, Nancy Logue, to treat [her]$_j$ speech impediment.

The most frequently addressed subtask of ACR is the pronoun resolution, which entails identifying the antecedents for anaphoric pronouns, particularly third person pronouns. A brief discussion of ACR systems, which are occasionally developed to address a subproblem like pronoun resolution, is presented in the sections that follow.

2.1.2.1 A Brief History

There have been periods in the history of ACR that are comparable to the history of most other NLP tasks. While the early efforts on ACR rely on handcrafted rules and heuristics, with the introduction of sizeable datasets in the 90s, research in the field also focused on development of learning based models. However, rule-based models were still active (e.g., Lee et al. [2013]) until recently. End-to-end neural models dominate the area in recent years due to the availability of more computational power and sizable data resources.

In this section, we give a brief historical development of ACR by referencing some of the key works in the field. We refer to Mitkov [2002] for a thorough summary of the literature on anaphora resolution[10] and its subtasks from the beginning to that date. Ng [2010, 2017] offers a concise history of machine learning models in coreference resolution. For the new developments in neural models, we refer to Sukthanker et al. [2020], Stylianou and Vlahavas [2021].

Early computational works on the coreference resolution task rely on handcrafted heuristics. For instance, Bobrow [1964] matches the textual patterns to resolve the anaphor-antecedent pairs, employing heuristics such as 'phrases with "this" refer to similar

[9] The numbers are computed from the statistics given in Table 1 in Poesio et al. [2018]

[10] The book mostly covers the works on coreferential anaphoric relations.

phrases (e.g., “This price” may refer to “The price” in the preceding sentence)’. Winograd [1972] uses more complex heuristics based on the linguistic properties of candidate antecedents; for instance, subjects are favored over objects. Later research benefits from more complex knowledge resources. For instance, Hobbs [1978] proposes a syntax-based pronoun resolution algorithm, known as the Naïve algorithm, that traverses parse trees in a particular order and chooses the first noun phrase with compatible features with the antecedent, such as assessing the agreement in number and gender. Lappin and Leass [1994] describe an algorithm for resolving third person pronouns that is primarily based on the salience information derived from syntactic parse-trees of the sentences in the text. The system uses procedures for identifying pleonastic pronouns and for computing the salience measures using certain parameters, such as grammatical role of noun phrases and recency (i.e., sentential distance between the pronoun-antecedent pairs). In addition to these mostly syntax-based approaches, in a different line of research, discourse information, especially centering theory [Grosz et al., 1995], has been used for automatic coreference resolution, for instance by Brennan et al. [1987]. Their research has been followed by the works of Strube and Hahn [1996], Tetreault [2001]. In a similar direction, Kehler and Rohde [2013] show how centering can be integrated with coherence-driven theories of pronoun interpretation.

The use of learning-based techniques became increasingly popular with the creation of coreference-annotated MUC corpora [Grishman and Sundheim, 1996, Hirschman and Chinchor, 1998]. In their influential work, Soon et al. [2001] implemented a supervised learning-based system that employs automatically extracted features from the text, such as the textual distance between a mention and a candidate antecedent, the gender, number, and semantic class of noun phrases. Their learning-based resolver showed a competitive performance with the existing, mostly rule-based systems on the MUC corpora Soon et al. [2001] and therefore encouraged research on learning-based models in this field. A series of machine learning models followed on the task over the next 15 years, e.g., Cardie and Wagstaff [1999], Ng and Cardie [2002], or sub-tasks of ACR, e.g., Bean and Riloff [2004], Bergsma and Lin [2006]. Most of these approaches were using a pipeline composed of several automatic NLP tasks such as Part-of-Speech tagging, parsing and noun phrase identification. The first end-to-end model is a neural model implemented by Lee et al. [2017] which instead of following a pipeline approach addressed the complete coreference resolution problem with a single neural network implementation. The model became state-of-the art when it was published, and inspired the ACR research of the following years, such as the works of Lee et al. [2018], Joshi et al. [2020].

As we briefly mentioned above in Section 2.1.1, CoNLL-12 shared task data, which has a clear separation of training, development and test sets, has become a standard benchmark for comparing the performance of newly developed systems with respect to the previous results. Metrics in the field introduced for evaluating the performance of automatic coreference resolution systems vary as we briefly present in Section 2.1.2.3. The performance scores indicated by these metrics do not always align and each metric garnered criticisms for various reasons [Recasens and Hovy, 2011, Moosavi and Strube, 2016]. In the CoNLL-12 shared task, the applied evaluation method combines three metrics, MUC, B-cubed and CEAF, often referred to as the CoNLL F1 score. Luo et al. [2014] introduces a reference implementation for the computation of the CoNLL F1 score which, similar to the data benchmark, is currently the standard evaluation method for comparing the performance of ACR systems. As a result, most of the recent learning based systems are evaluated on the CoNLL shared task dataset by using the CoNLL F1 score.

2.1.2.2 Coreference Resolution Models

Computational research on ACR is dominated by three major learning based models. In this section we briefly examine these models.

Mention-pair models Soon et al. [2001] implemented a learning based model that was motivated by the work of Aone and William [1995]. The coreference resolution model they applied is known as the **mention-pair model** and is based on a trained binary classifier that determines whether or not two mentions are coreferential. A separate clustering mechanism is employed to combine the pairwise decisions into coreference chains. In order to identify the mention-antecedent pairs, Soon et al. [2001] employed 12 features including the distance between a mention and a candidate antecedent as well as the gender, number, and semantic class (e.g., person, organization, date, time) of noun phrases. Following the work of Soon et al. [2001], the mention-pair model has been extensively used, for instance in the works of Ng and Cardie [2002], Ponzetto and Strube [2006], Uryupina and Moschitti [2015].

There are two main areas where mention-pair models are considered to be weak. First, each preceding mention for a given mention is evaluated individually as a potential candidate for an antecedent. As a result, it is impossible to compare a potential antecedent's performance to that of the others. In addition, mention-pair models are criticized for having limited **expressiveness** (i.e., they can only employ local features defined on an anaphor and a candidate antecedent) Ng [2017]. This becomes problematic particularly when the candidate antecedent is not sufficiently informative, as in the case of a pronoun. These flaws have been addressed by two separate lines of research as we present below.

Mention-ranking Models Rather than taking each candidate separately, **mention-ranking models** rank the preceding mentions as potential antecedents for a given mention and choose the highest scoring mention as the antecedent. Therefore, compared to mention-pair models, a better antecedent selection mechanism is offered by the mention-ranking approach. Although this approach, similar to the mention-pair model, is limited in terms of expressiveness, it is efficient and fast to apply in either rule-based or learning-based models. Therefore, it is highly exploited in the field, for instance in the works of Denis and Baldridge [2008], Wiseman et al. [2015].

Entity-based Models Another common approach in ACR research is called the entity-mention approach. It addresses the second weakness of the mention-pair models (i.e., **expressiveness**). Entity-mention models consider the candidate antecedents not in isolation but in the partially constructed coreference clusters they belong to. An entity-mention classifier determines whether a mention belongs to a preceding coreference cluster or not. These models can extract more information than mention-pair models, and therefore they are more expressive, even if some of the antecedent mentions are uninformative (e.g., a pronoun) because they can employ features from all members of a partially built entity cluster. Luo et al. [2004], Yang et al. [2008] are example models that implement the entity-mention approach. In order to combine the power of mention-ranking and entity-mention models, Rahman and Ng [2009] merge these models and call it a cluster-ranking model, that is further applied in various systems such as by Yu et al. [2020].

2.1.2.3 Evaluation Metrics

Several metrics have been introduced for evaluating the experimental performance of coreference resolution systems. Coreference resolution can be considered as a task with two main aspects: detection of mentions and establishment of coreference links between the mentions (i.e., composition of entity representations). The introduced metrics vary in terms of which aspect of the coreference resolution task they focus on. Below we give a brief overview of the metrics currently in use.

MUC The MUC is a link-based metric [Vilain et al., 1995] that was developed to evaluate the systems competing in MUC-6 and MUC-7 coreference resolution tasks. The precision and the recall are computed by dividing the number of common links between gold and predicted mentions by the number of links in predicted and gold chains, respectively. MUC disregards the singleton chains, because they don't include links; the approach received criticism for that reason.

B^3 The B^3, also known as B-cube or B-cubed, is a mention-based metric [Bagga and Baldwin, 1998]. Precision and recall scores are computed for each mention separately and then the weighted sum of these scores were used to compute the precision and the recall for the entire data. B^3 is criticised for assuming that the predicted and the gold mentions are the same, and therefore, not being able to handle spurious and missing mentions in the predicted set.

CEAF The CEAF Luo [2005] is computed by first making a one-to-one mapping of the predicted entities to gold entities. Then the recall and the precision are computed based on the similarity of the mapped entities (i.e., coreference chains), which is measured according to a predefined similarity metric. Different versions of the CEAF exist in the literature that vary according to the method used for similarity measuring. For instance, CEAF$_\mathrm{m}$ computes the similarity of the entities using the number of common mentions between these entities, hence it is a mention-based metric.

BLANC The BLANC [Recasens and Hovy, 2011, Luo et al., 2014] is a link-based metric. Different from the MUC, the BLANC not only considers the coreferential links in the predicted and the gold entities but in addition considers the non-coreferential items in computing the precision and the recall. For instance if a,b,c,d are four mentions and the a,b and the c,d are representing two separate coreference chains, the links between a/b and c/d are counted as coreferential links. Furthermore, the non-coreference links a/c, a/d, b/c, and b/d are additionally considered in the evaluation. The BLANC score is the average of coreferential and non-coreferential F scores.

LEA LEA [Moosavi and Strube, 2016] is a link-based entity aware metric. Moosavi and Strube [2016] argue that B^3, CEAF and BLANC are not reliable for recall-precision analysis. They describe scenarios in which they make several changes in the output of a system which, in theory, should not change the precision or recall, depending on the change. However, they observed that these theoretical assumptions were not satisfied by the above-mentioned metrics. Therefore, they propose LEA as a new coreference resolution evaluation metric which, they consider, do not carry the shortcomings of the other metrics. LEA takes the importance of the entities (i.e., the longer the coreference chain, the more important the entity) and computes the precision and the recall by using this value and the score that is computed based on the correctly identified coreference links.

CoNLL Score As we briefly discussed above, there exist various metric proposals for assessing the empirical outcome of the automatic coreference resolution systems. In the CoNLL-2011/2012 shared tasks [Pradhan et al., 2012], the evaluation metric used for

[10] According to the number of mentions in the chain that the mention in concern belongs to

ranking the participating systems was the average of MUC, B^3 and CEAF. A reference implementation for this metric was distributed by the organizers which was later extended to include the BLANC score as well [Pradhan et al., 2014]. This metric, often called the CoNLL score, was heavily used to report the performance rates by the automatic systems developed since then and has become the most common standard for comparing the outcome of different systems.

2.1.2.4 Current Discussions

The state-of-the-art ACR models, e.g., [Xu and Choi, 2020, Wu et al., 2020], currently perform with above F1 score of 80% on the CoNLL-12 benchmark data. On other domains or datasets, however, the success rates drop dramatically. Developing ACR systems which can run with high performance scores on different topic domains (e.g., biomedical domain [Lu and Poesio, 2021]) and genres (e.g., dialogs [Khosla et al., 2021], literary documents [Bamman et al., 2020]) beyond OntoNotes has become a subject of increasing interest in recent years (for a detailed review of this line of research, see Chapter 6).

The vast majority of ACR models and the current large corpora (OntoNotes, MUC, ACE) account for the nominal coreference. The community's attention has started to turn in recent years to more sophisticated cases of anaphora that are not contained in OntoNotes, like bridging anaphora and discourse deixis through CRAC-2020 and 2021 shared tasks [Ogrodniczuk et al., 2020, 2021] as well as coreference in understudied lowersource languages in CORBON shared tasks [Ogrodniczuk and Ng, 2016, 2017].

As briefly explained in Section 2.1.1, existing coreference-annotated corpora have various annotation schemes and data formats, which makes using these corpora together difficult. A recent initiative called Universal Anaphora Initiative[11] has been launched by some researchers in the field with the goal of creating a universal annotation scheme and harmonizing the existing corpora accordingly. In the same vein, a multilingual dataset is created for the CRAC 2022 Shared Task[12] that includes texts from different languages that are annotated using a uniform scheme.

Although there are improvements in the performance rates for the entity coreference, it is still not a solved problem. For instance, the sentence pair in 2.8 and 2.9, taken from [Levesque et al., 2012, p.559] illustrate one challenging setting, first introduced by Terry Winograd [Winograd, 1972] and therefore called the Winograd schema. In Winograd schema, only one or two words are different in two sentences that otherwise agree in terms of syntactic structure and tokens. The resolution of the pronouns is impacted by the existing discrepancies in the sentences. For instance, in order to resolve the pronoun "she" in the sentences 2.8 and 2.9, more background knowledge that is not expressed in the words of the sentence is needed. Levesque et al. [2012] associates the use of background knowledge in these cases with the act of **thinking** and proposes a test as an alternative to the Turing test to evaluate whether a machine exhibits a human-level intelligence; this is called the Winograd Schema Challenge. Since the challenge was proposed, competitions encourage the researchers work on the challenges of the cases of the Winograd Schema instances.

(2.8) $[\text{Anna}]_i$ did a lot better than her good friend $[\text{Lucy}]_j$ on the test because $[\text{she}]_i$ had studied so hard.

(2.9) $[\text{Anna}]_i$ did a lot worse than her good friend $[\text{Lucy}]_j$ on the test because $[\text{she}]_j$ had studied so hard.

[11] https://github.com/UniversalAnaphora/UniversalAnaphora

[12] https://www.aclweb.org/portal/content/5th-workshop-computational-models-reference-anaphora-and-coreference

Addressing the gender bias in pronoun resolution is a topic of contemporary research, which is especially important for the languages with gendered pronouns such as English. The term “gender bias” refers to the propensity for repeating gender stereotypes about the gender of people in particular roles, such as the inclination to refer to “the surgeon” as “he” and “the secretary” as “she”. Rudinger et al. [2018] shows that gender bias is systematic in the off-the-shelf coreference resolvers they tested. Studies addressing this bias, e.g., by Cao and Daumé [2021], are recently emerging in the field.

2.2 Twitter

Twitter is a social media platform launched in 2006. It is known as the most popular **microblogging**[13] service, with over 330 million monthly active users, by the start of 2019[14].

Twitter users are spread all over the world and post in many different languages. As shown in Figure 2.1 the most popular language used in Twitter messages is English (32%), followed by Japanese, Spanish, Korean, Arabic, Portuguese, Thai, Turkish and French.

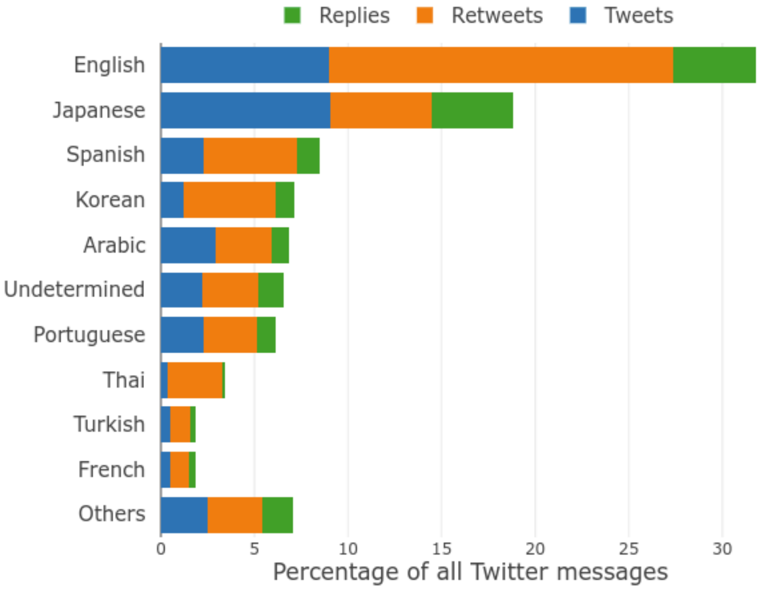


Figure 2.1: Most popular languages on Twitter (Data taken from `https://www.vicinitas.io`)

2.2.1 Usage

Twitter lets anyone read messages without having to sign up. However, in order to post messages on Twitter (also known as **tweets**), one must first sign up for the service and create an **account** on the platform. The act of posting in Twitter is also known as **tweeting**. All registered users have **usernames**, which are unique identifiers that always begin with the @ symbol. A tweet may contain photos, videos, links, emojis and text. As a result, they are multi-modal. Users frequently use **hashtags**, which are keywords or key phrases preceded by the '#' symbol, to identify the topic of a tweet. Hashtags connect conversations and make it easier to find tweets about the same topic. Users **follow** the

[13] broadcasting **short** messages to the other subscribers of the service

[14] https://en.wikipedia.org/wiki/Twitter

[14] `https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets#language`

accounts and topics they are interested in to see the most recent tweets by those accounts or on those topics. The main page for users is called a **timeline** that displays a stream of tweets from accounts and topics a user has chosen to follow on Twitter.

Twitter users send tweets to spread information and to interact with other users. In order to make a post on Twitter, users can either create a new tweet, broadcast an existing one (i.e., **re-tweet**) or reply to an existing tweet. User interactions through the reply tweets build conversations. Conversations on Twitter are organized into **tree** structures, as illustrated in Figure 2.2[15]. Users can reply at every level of the conversation tree. The usernames of the conversation participants in a conversation structure, introduced by the @ sign, are automatically added to the content of the reply message in Twitter.

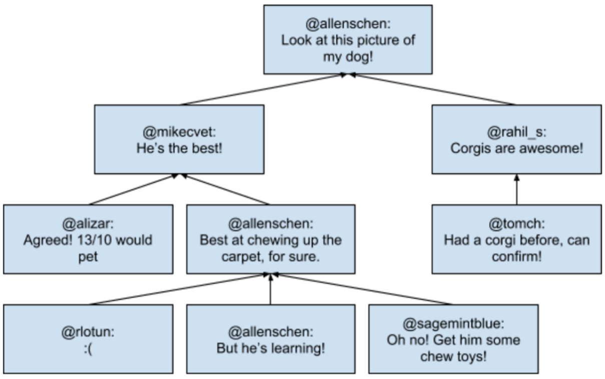


Figure 2.2: A conversation tree in Twitter

2.2.2 Language in tweets

Unlike most other popular social media platforms, Twitter imposes a character limit on tweets. The restriction was originally set at 140 characters, but in 2017 Twitter decided to raise it to 280 characters. Because of this limitation, users need to be creative with their language in order to fully convey their ideas. It emphasizes character-reducing writing style such as use of acronyms, abbreviations, emojis to express the emotions and shortened variations of the words [Boot et al., 2019].

As soon as a tweet is accessible, users can read and reply to it whenever they want. Therefore, Twitter interactions are in theory asynchronous. The stream of tweets on users' timelines, however, typically changes so quickly that Twitter interactions become practically near-synchronous; otherwise, users risk missing the tweet to which they wish to answer [Scheffler, 2017]. This interactive unplanned nature of the platform fosters the development of a community that uses creative, genre-specific terminology and phrases [Zhang et al., 2019]. For instance, Liu et al. [2011] found over 4 million unique out-of-vocabulary words in the English tweets in the Edinburgh Twitter corpus [Petrović et al., 2010]. For example, the words 2gether, togetha, t0gether, and tgthr are out-of-vocabulary variations originated from the word "together" in the Edinburgh Twitter corpus.

In summary, the characteristics of Twitter as a platform influence the language that is employed. Below, we provide examples of tweets that include peculiarities/non-standard usages in Twitter language such as deliberate or unintentional misspelling, non-standard punctuation, capitalization, abbreviations, vocabulary, and syntax. These examples are taken from [Eisenstein, 2013, p. 359].

(2.10) Work on farm Fri. Burning piles of brush WindyFire got out of control. Thank

[15] Taken from the blog `https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/building-and-serving-conversations-on-twitter`

God for good naber He help get undr control PantsBurnLegWound. (Senator Charles Grassley)

(2.11) Boom! Ya ur website suxx bro (Sarah Silverman)

(2.12) ...dats why pluto is pluto it can neva b a star (Shaquille O'Neil)

(2.13) michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her. (Ozzie Guillen)

2.2.3 Twitter for data analysis

Through its API[16] (**A**pplication **P**rogramming **I**nterfaces), Twitter provides programmatic access to its data to users, businesses, and developers. It is possible to build a corpus to work on by downloading the tweets for specific languages, dates, topics, etc. through that API.

Twitter users publish messages in real time on a range of subjects, such as current events (e.g., the COVID outbreak), opinions about political issues, complaints, positive and negative feedback for products they use, and many more topics. Therefore, together with its growing popularity, Twitter has become an interesting dataset for studies investigating general human behavior and the linguistic features of the texts. Businesses examine Twitter data to determine, for example, how their target markets will respond to their products [Jansen et al., 2009]. Among the many other types of computational applications, we can mention research with practical implications such as sentiment analysis[Agarwal et al., 2011, Nakov et al., 2016], argument mining[Dusmanu et al., 2017, Schaefer and Stede, 2021], and named entity recognition [Ritter et al., 2011, Pipitone et al., 2017] which work on Twitter data for commercial and scientific interests. Twitter data is also interesting for empirical linguistic analysis such as corpus-based research on language change [Xu, 2017, Donoso and Sánchez, 2017, Dijkstra et al., 2021]. To the best of our knowledge, Twitter data has not yet been explored for the empirical examination of linguistic aspects of coreference and computational aspects of coreference resolution apart from our own work Aktaş et al. [2018, 2020], Aktaş and Kohnert [2020], Aktaş and Stede [2020]. In this work, we contribute to the body of scientific research on this genre in this way.

2.3 Exploring Variation in Language

2.3.1 Clarification of Terminology

The fundamental premise of research on linguistic variation is that the diversity we can observe in language is not random. As clarified by [Biber and Conrad, 2009, p.4]:

> At the highest level, linguistic variation is realized as different languages (e.g., Korean, French, Swahili). At the lowest level, linguistic variation is realized as the differences between one speaker compared to another speaker, or as the differences between two texts produced by the same speaker.

We are interested here in the linguistic variation that happens between these levels in different text[17] varieties produced in a language. Text varieties are widely described in the literature using the terms *register* and *genre*. However, these terms are employed in various ways throughout the literature and there is no widespread agreement over their

[16] http://dev.twitter.com

[17] We use the term **text** very broadly here, in the sense of "any passage (of language), spoken or written, of whatever length, that does form a unified whole" [Halliday and Hasan, 1976, p.1]

usage. In this section, we define the terms we use in this work without attempting to reconcile their various existing usages. For thorough reviews of the terminology use in linguistic variation research, we refer to Lee [2001], Biber and Conrad [2009].

In our work, the term **mode** distinguishes the language production means broadly between spoken, written and signed. **Medium** or channel is the physical means of text realization (e.g., telephone, Twitter platform). The terms register and genre have been extensively used in the literature to describe varieties of texts that differ, for instance, in terms of situational or functional characteristics. Although they are frequently used interchangeably, there are also approaches that clearly distinguish between them. For instance, within Systemic Functional Linguistics (SFL), Martin [1985] defines "genre" and "register" as two different semiotic planes. A genre is "a staged, goal-oriented, purposeful activity in which speakers act as members of [a] culture" (cited in [Jonsson, 2016, p.36] from Martin [2001]). On the other hand, "register is defined as a particular configuration of field, tenor, and mode[18] choices (in Hallidayan grammatical terms), in other words, a language variety functionally associated with particular contextual or situational parameters of variation and defined by its linguistic characteristics." [Lee, 2001, p.42] (Footnote added by the author of this thesis.) [Jonsson, 2016, p.36] summarizes the discussion on register and genre in SFL as follows:

> Genre thus corresponds roughly to "context of culture" and register to "context of situation".

Biber and Conrad [2009] also make a distinction between register and genre, which has conceptual similarities with SFL, particularly with the genre perspective emphasizing the conventional features:

> The underlying assumption of the register perspective is that core linguistic features (e.g., pronouns and verbs) serve communicative functions. As a result, some linguistic features are common in a register because they are functionally adapted to the communicative purposes and situational contexts of texts from that register. In contrast, the genre perspective focuses on the conventional structures used to construct a complete text within the variety (for example, the conventional way in which a letter begins and ends) [Biber and Conrad, 2009, p.2].

In contrast to these approaches, many of the corpus-based studies do not make a distinction between these concepts and instead adopt one of them and disregard the other in referring to the text varieties they examine. For instance, while Fox [1987], Toole [1996], Lee [2001] use the term "genre" to address the text varieties they investigate, Biber [1992], Kunz and Lapshinova-Koltunski [2015] use the term "register", without a substantial justification of the terminology. Text groupings in big corpora such as in OntoNotes [Weischedel et al., 2013] and ARRAU [Uryupina et al., 2016] are also addressed with the term "genre". In our work, we take a theory-agnostic approach towards the text varieties we investigate. For the purposes of this study, following the terminology used in the work of Fox [1987] and the categorization used in the OntoNotes corpus, we adopt the term **genre** to address the collections of texts compiled for the corpora we examine. Genres, in our case, are characterized with mode (spoken vs written) and medium (e.g., telephone, TV, radio, newspaper, digital media). The scope of granularity to account for the differences between text varieties vary, even with a relatively broad definition of genre. For instance, it is a matter of choice to consider "the telephone conversations" and "the broadcast conversations" as separate genres or sub-genres of "conversation". In

[18] The terms *field*, *tenor* and *mode* roughly mean the type of social action, the relationships between participants, and the channel of communication.

our case, we rely on the distinctions made in OntoNotes and treat all separate sections of OntoNotes as a distinct genre (e.g., telephone conversations, broadcast conversations, broadcast news, news) as they differ at least in terms of either mode or medium.

The term "domain" also has at least two uses. The narrow one is for referring to the topic in the text (e.g., biomedicals, economics, sports). We employ the phrase "topic domain" for referring to that narrow use. We also sometimes use "domain" in a very broad, nonspecific sense, following the common convention in NLP. In this sense, we refer to linguistic data with particular features that vary depending on the context as part of a **domain**. A domain can be, for instance, a dataset or a specific genre.

2.3.2 Spoken and Written English

The question of how speech and writing relate to one another has had the attention of linguists for a long while. Exactly one hundred years ago, [Woolbert, 1922, p.271] emphasized the need for examining the similarities and divergences in spoken and written language: "Speaking and writing are alike – and different. Just how like and how different has never been adequately stated." A great number of publications regarding the empirical investigation of spoken and written language followed Woolbert's paper. For a comprehensive survey of the literature in this area, we refer to the review of Jonsson [2016].

Examining the grammatical and lexical differences between spoken and written language has been a major area of research from the early comparative studies on mode contrasts. The list of documented variations in the spoken-written dichotomy includes the following:

- Writing has greater vocabulary variety (e.g., Chafe and Danielewicz [1987], Biber [1988])
- Writing has higher lexical density (i.e., a higher ratio of content words) (e.g., Halliday [1985, 2004])
- Speaking includes more demonstrative pronouns and deictic terms (e.g., Ochs [1979], Chafe and Danielewicz [1987], Biber et al. [1999])
- Speaking includes more personal pronouns (e.g., DeVito [1966], Chafe [1982])
- Writing includes longer words (e.g., Zipf [1949])
- Writing includes longer sentences (e.g., Ochs [1979], Chafe [1982])
- Writing includes more relative clauses (e.g., Ochs [1979], Chafe [1982])
- Speaking includes longer textual distance between anaphoric expressions and their antecedents [Fox, 1987]

Several researchers have argued that there exist underlying binary dimensions concerning the spoken-written distinction. For instance, while Ochs [1979] explains the distinctions between spoken-written modes in terms of planned versus unplanned dichotomy, Lakoff [1982] explores the same dimension addressing the contrast between spontaneous versus non-spontaneous discourse. Chafe [1982] accounts for the distinctions in terms of two dimensions, namely, fragmentation versus integration, and involvement versus detachment. Chafe [1982] identifies sets of co-occurring features that characterize written and spoken language. In that framework, written language is more integrated (e.g., having more frequent use of relative clauses) but is more detached in terms of the relationship between the writer and the audience. Spoken language is more fragmented (e.g., sentences frequently being interrupted by pauses) but has more involvement among interlocutors (e.g., having frequent use of first and second person pronouns and other deictic terms). Tannen [1982b] explores these dimensions in the spoken and written narratives, arguing that they can coexist in a single text:

> [..] features of integration and involvement, which Chafe finds characteristic of writing and speaking respectively, can be combined in a single discourse type. [Tannen, 1982b, p.2]

In light of her findings, Tannen [1982a] argues that it is helpful to think of the spoken-written disparities as a continuum (oral vs literate continuum, in her model), rather than mutually exclusive modes of language production. Koch and Oesterreicher [2012][19] share this approach that spoken and written language are not clear-cut notions and introduce a two-dimensional model based on the German language to capture the spoken vs written modes as communication mediums on the one hand, language of immediacy (conceptually oral) and language of distance (conceptually written) forms on the other hand. The distinction in the conceptual dimension is gradual. The conceptually written pole is characterized, for instance, by grammatically complete and complex clauses, whereas conceptual orality includes fewer organized clauses that frequently lack temporal planning and are occasionally incomplete. The diagram in Figure 2.3 shows different genres (forms of discourse in Koch and Oesterreicher's terminology) in order of decreasing 'spoken' and increasing 'written' conception. According to this model, a text may be conceptually spoken or conceptually written regardless of the production medium (i.e., graphic or phonic form).

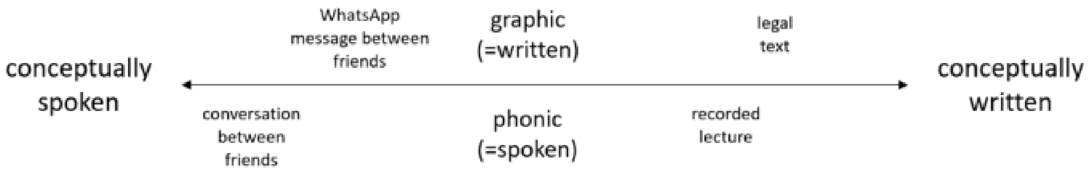


Figure 2.3: Simplified version of [Koch and Oesterreicher, 2012, p.444]'s model (Taken from [Leuckert and Buschfeld, 2021, p.8])

In a similar manner, [Biber, 1988, p.199] states that:

> there is no single, absolute difference between speech and writing in English; rather there are several dimensions of variation, and particular types of speech and writing are more or less similar with respect to each dimension.

Using the findings of previous research in the field, Biber [1988] identifies 67 features as being likely to distinguish between speech and writing, including the frequency of tense and aspect markers, nominal forms, and coordinating clauses. Biber reduces these large number of features to a small set of factors by detecting the co-occurring features. To identify the features that frequently co-occurred in texts, Biber employs a statistical method known as factor analysis. The groups of co-occurring features address the following six dimensions of linguistic variation in Biber's approach, and as a result it is known as the multi-dimensional approach. The multi-dimensional approach indicates that a text can represent any of the dimensions regardless of its production mode.

[19] This is the English translation of the article by Peter Koch and Wulf Oesterreicher, entitled "Sprache der Nähe—Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte," that was published in German in 1985.

- Dimension 1: Involved vs Informational production
- Dimension 2: Narrative vs Non-narrative concerns
- Dimension 3: Explicit vs Situation-dependent reference
- Dimension 4: Overt expression of persuasion
- Dimension 5: Abstract vs Non-abstract information
- Dimension 6: On-line informational elaboration

The continuum approach to the spoken-written distinction in general, as well as Koch and Oesterreicher [1985, 2012]'s two-dimensional and Biber [1988]'s multi-dimensional models encouraged the research examining the variation between spoken-written modes to focus on distinctions between genres. The analyses therefore included a variety of genres, and the results discussed how these genres related to one another in terms of the variables that were being examined.

Recent studies on genre-based distinctions examine the digital texts that have emerged as a result of the widespread usage of computer-mediated communication (CMC) during the past 30 years. For instance, Jonsson [2016] makes a multidimensional analysis of English computer-mediated conversational writing represented by "Internet relay chat" and "split-window ICQ chat" that are identified by the author as instances of synchronous CMC and supersynchronous CMC, respectively. The study indicates that the dimension scores of the conversational writing genres that are computed with respect to the features introduced by Biber [1988] reveal that in the majority of these dimensions conversational writings are closer to oral conversations. Jonsson also provides a thorough linguistic analysis of the conversational writing genres included in the study. We refer to the actual study for further discussions of the analysis.

Although Koch and Oesterreicher [1985, 2012] was a well-known model in German linguistics, it did not receive much interest from the English speaking community until recently. Schaefer [2021] provides a detailed analysis of the reasons for the model's "non-reception" in English-speaking linguistics. A special edition of the Anglistik journal featuring the model's application to English texts was published in 2021[20]. In that issue, Ronan [2021] examines the tweets from Donald Trump's Twitter account, @realDonaldTrump, in the context of Koch and Oesterreicher [2012]'s two-dimensional model. Her analysis demonstrates how Trump's tweets frequently employ aspects of the language of immediacy, such as spontaneity and proximity, bringing his tweets closer to the spoken end of the continuum. In another article in the same issue, Rüdiger [2021] investigates the speech in eating shows on Youtube. The eating shows are described by the author as public performances of food consumption that are watched asynchronously by the audience. The analysis indicates that "despite being produced primarily in the spoken mode, eating shows constitute a fascinating mix of characteristics associated with immediacy and distance" [Rüdiger, 2021, p.1].

2.3.3 Variation in Coreference Patterns

Studies included in this section use quantitative corpus-based methods to examine the variation of referential phenomena through relevant linguistic features. Variation can be explored in different domains such as across languages [Lapshinova-Koltunski, 2015, Kunz and Lapshinova-Koltunski, 2015, Engell, 2016], across regional language varieties [Neumann and Fest, 2016], across production modes (spoken, written) [Fox, 1987, Biber,

[20] `https://angl.winter-verlag.de/issue/ANGL/2021/2`

1992, Amoia et al., 2012] and across genres [Swanson, 2003]. We include the quantitative results and discussions of these studies related to variation of referential strategies in English with respect to genre and production mode.

Kunz [2007] proposes a set of linguistic features for an empirical examination of coreference variation in her in-depth review. The following features are considered for the comparative study of nominal coreference:

- Distribution of referring expressions
  - Number of coreference chains
  - Length of chains (i.e., number of referring expressions in one coreference chain)
  - Distance between coreferential expressions
- Linguistic structure of referring expressions
  - Form of referring expressions (e.g., definiteness markers, plural/singular nouns)
  - Type of relation between two anaphoric expressions (i.e., identity, repetition, hyponymy, hyperonymy, synonymy)
  - Syntactic role of referring expressions (e.g., subject, object)
  - Syntactic position of referring expressions (e.g., preverbal, postverbal)
- Cognitive status of referents
  - Information status of referent (e.g., activated, new)
  - Type of reference (i.e., reference to a unique/specific/generic referent)

The comparative studies covered in this section use mostly the descriptive frequency-based statistics of the subset of the linguistically motivated criteria given above. Some of the studies include extra features that are not on this list, which we discuss separately in the overview.

In her seminal work, **Fox [1987]** examines the use of anaphoric third person singular human references (pronouns and noun phrases) in spoken and written texts. The data is composed of spoken conversations (face-to-face and telephone conversations) and written expository prose (newspaper articles and psychoanalytic biography). Fox makes the statistical description of the anaphoric patterns in terms of two features:

- Distance between referring expressions and their antecedents (in terms of clauses)
- Frequency of syntactic categories of referring expressions

Her results indicate that conversations allow longer clausal distance than written texts between pronouns and their antecedents (i.e., 2.52 clauses in conversations vs 1.21 in written texts). In addition, referential NPs are more frequent in expository texts than in conversations (i.e., 47% in written texts vs 22% in conversations).

Fox first discusses her results referring to the topic continuity model [Givón, 1983] which Fox calls the "traditional theory of anaphora". The main premise of the model is that the form of an anaphoric expression is determined by topic continuity. The continuity of topic can be primarily assessed by using three factors: the distance (clausal or sentential), the number of interfering referents between a referring expression and its antecedent, and thematic continuity (i.e., whether the referring expression is paragraph-initial or paragraph-medial). The model associates the use of pronouns as referential expressions with high continuity of topic and the use of NPs with low continuity or discontinuity of topic. Fox checks the implications of this model in her data. Examination of the distance feature for pronouns and NPs indicate that the expected correlation between the distance and the referential form is not observed for an important portion of the data. Fox discusses instances in which pronouns are used instead of NPs where a distance based model, such as the topic continuity model, would have anticipated the use of NPs (i.e., when there is a great distance between the referring expressions) or where NPs are used instead of pronouns when a distance based model would predict the use of pronouns (i.e., when there is a very short textual distance between the referring expressions). As a result, Fox proposes considering the anaphoric distribution in discourse within a discourse structure framework because the problem cannot be explained only in terms of textual linearity.

Fox examines spoken and written texts using distinct discourse frameworks, namely conversational analysis [Sacks et al., 1974] for addressing the structural characteristics in spoken data and RST [Mann and Thompson, 1988][21] for written texts. The structural units of conversational analysis are adjacency pairs. In the simplest form of adjacency pair there exist two utterances where the first one initiates a situation for which the second utterance is relevant. Typical examples of adjacency pairs are question-answer, offer-acceptance, and invitation-acceptance. The use of separate structural approaches for different modes is grounded by addressing the differences in situational characteristics of the examined genres.

Fox looks into referential choice in different environmental settings with respect to the existence of interfering referents. The frequency of the referential pronouns in these environments is presented in Table 2.1. Her findings suggest that pronoun usage decreases for both modes if there exist interfering referents. Investigation of patterns in same-gender environment (i.e., the presence of same-gender referents) indicates that in written texts, pronouns are used in the same-gender environment only if the referent is the subject of the proposition it has expressed. For conversational data, pronoun usage in the same-gender environment is not as constrained as in the written texts, but there still exist some restrictions on the choice of the referential expressions which are mostly attributed to the structural embedding in terms of adjacency pairs in Fox's analysis. Fox argues that a significant portion of the data can only be explained by structural means. For instance, two textually distant referential expressions can in fact be close in the hierarchical discourse structure, which can explain the use of pronouns when there is a long linear textual distance with the antecedent. Therefore, discourse structure is a key component in explaining the choice of referential expressions.

Pronouns	No interfering referent	different-gender int. ref.	same-gender int. ref.
Written texts	64%	39%	13%
Spoken texts	94%	72%	57%

Table 2.1: Pronoun usage in different environmental settings (adopted from Table 6.6 in [Fox, 1987, p.148])

Biber [1992] analyzes the use of anaphoric referential expressions in nine written and spoken genres. The written genres included in the study are press spot news reportage, legal documents, humanities and technical academic prose, and general fiction while the spoken genres are face-to-face conversations, sports broadcasts, government spontaneous speech, and sermons. Biber identifies and classifies all referring expressions in the data with a semi-automated procedure (i.e., manual post-editing of automatically annotated data). The study examines the following linguistic features and finds the results presented below.

[21] Fox cites the earlier versions of this publication.

- **Distribution of referring expressions**:
  - **Number of referring expressions**: Spoken genres tend to have more referring expressions than written genres.
  - **Number of referents** (including singleton chains): The two written genres of academic writing have two of the highest numbers of different referents while conversations have the lowest number of different referents.
  - **Length of coreference chains**[22]: Conversations have the longest chains (*mean* 2.91), whereas two academic prose genres have the shortest chains (*mean* 1.40 and 1.58). Spoken genres seem to have longer chains than written genres with an exceptional written case of general fiction which is closer to spoken genres than to written genres in this respect.

- **Distance between referring expressions** (in terms of the number of interfering references): The average distance is longer in written genres than in spoken genres with the exceptional case of sports broadcasts which is closer to the written genres than spoken genres in this respect.

- **Information status of a referent** (*given* vs *new*): Conversations and broadcasts tend to have more *given* entities than *new* entities whereas the two written genres of press reportage and academic prose have the opposite tendency. However, the three genres (i.e., speeches, sermons, and fiction) from both written and spoken modes share a tendency to include a roughly equal number of given and new referring terms.

- **Type of referring expressions**: Pronouns are used more frequently than lexical repetitions in spoken genres, with the exception of broadcasts, which use repetitions more frequently. With the exception of fiction, which uses pronouns more frequently, written genres have a tendency to use lexical repetitions more frequently than they do pronouns. Biber examined the frequency distributions of pronouns vs nouns across different genres (conversations, fiction, news and academic prose) in Biber et al. [1999, p. 235]. The findings of Biber et al. [1999] confirm that pronouns are more common than nouns in conversations. At the other extreme, academic prose and news contain higher usage of nouns over pronouns.

- **The syntactic position of referential expressions** (for example, whether they are used in main clauses, relative clauses, prepositional clauses, etc.): The distinction of genres according to this feature is inconclusive. But one notable observation is that referential expressions are more frequent in prepositional phrases than in main clauses.

Biber discusses the implications of the quantitative results by referring to the situational characteristics of genres. As we noted above, spoken genres tend to have a higher number of referring expressions than do written texts. However, even though they contain a higher number of referring expressions, conversations tend to have a lower number of referred entities , and hence, longer coreference chains. According to Biber, this finding suggests that people prefer not to change the topic in conversations.

Biber uses exploratory factor analysis to compute the dimension scores for the referential features investigated according to the multi-dimensional methodology described in Biber [1988]. In his analysis, fiction and spoken genres of broadcasts, conversations,

[22] Biber analyzes the linguistic features in only the first 200 words, therefore individual numbers may not be representative. But the comparison of numbers for different genres can still be indicative for the differences and similarities of the genres.

and sermons are classified as having narrative characteristics. This finding is used as an argument to explain the similarities of the frequency distributions (e.g., type of referring expressions) of the linguistic features in the genres that originally belong to different language modes.

With a theoretical perspective, **Toole [1996]** argues that Accessibility Theory [Ariel, 1990] can explain the choice of referential forms across all genres. The descriptive analysis of eight texts from fiction novels, academic book reviews, informal conversations, and interviews provide the empirical data for the study. According to Ariel [1990], referential choice is driven by the degree of accessibility of the referred entity at the precise moment when the referring term appears in the discourse. For instance, pronouns show a high accessibility level while full names and lengthy descriptions show a low accessibility level for the referred entity. Ariel [1990] identifies four factors influencing the accessibility level:

- Distance between referring expressions and their antecedents(in terms of propositions[23])
- Number of interfering referents
- Topicality (i.e., whether the entity is the discourse topic)
- Unity (i.e., whether the antecedent is in the same discourse unit with the referring expression)

Ariel computes the magnitude of the accessibility level using the above-mentioned four features. For instance, for an entity A, if the last mention of that entity is in the same proposition with the referring expression under concern, 4 is added to the accessibility level of the entity (i.e., the contribution of distance); if there is a mention of another referent between the referring expression and its antecedent, the accessibility level is decreased by 1 (the contribution of interfering referents). Toole [1996] computes the accessibility levels for all the referring expressions in the dataset. The analysis indicates that there is a correlation between the referential form and their accessibility levels. Toole states that her observations confirm the predictions of Accessibility Theory regardless of the genre. For instance, approximately 72% of the pronouns refer to highly accessible entities and approximately 63% of long definite NPs refer to low accessibility entities. Toole [1996] argues that violations of the predictions of Accessibility Theory can be addressed by considering other potential factors influencing the accessibility levels, such as types of verbs in the propositions.

Swanson [2003] explores the idea of using coreference phenomena as a means of identifying genres by examining the qualitative and quantitative features of coreference in various texts. Academic journals, news magazines, and fiction narratives are the three written genres within and between which several aspects of coreference are studied. To explore intra-genre patterns, three works from each genre, written by three different authors, are included in the analysis. The common theme in all of the included texts is Middle Eastern politics. The first 30 sentences from each text are examined, and coreference relations are manually annotated. The study takes both identity and non-identity relations[24] between referring expressions into account. *Reiteration*, *Inclusion* and *Specification* are three relation types identified in the study. While the Inclusion and Specification categories cover hierarchical relations like being a part of, a type of, an instance of, etc., Reiteration captures identity relations.

[23] Proposition is chosen as the distance unit because it is considered to be applicable both spoken and written texts

[24] We refer to only identity relations as coreference relations, but different from us, Swanson calls all relations coreference relations. In examining their study, we'll stick with their decision.

Swanson first compares the frequency and distribution of investigated relations for all the texts and makes intra- and inter-genre analysis in terms of these measures:

- **Number of coreference relations**: The narrative texts have the highest average number of coreference links. In addition, while the texts belonging to the other two genres display uniform distribution in terms of the number of coreferential links, the narrative texts do not exhibit regularity in this respect (e.g., "1108 coreferential links in the first narrative text" vs "393 in the third"). Despite the inconsistency found in narrative texts, the author believes this characteristic to be a contender for a genre indicator due to the regularity found in academic journals and news magazines as well as the notable difference in the average number of relations between genres.
- **Number of links between any two sentences**: Compared to other genres, academic writings have a higher percentage of sentences with at least one coreference link to another sentence (i.e., "70% connected sentences in academic texts" vs "57% in narratives" vs "42% in news magazines"). The average number of links between sentences also displays differences between genres (i.e., "2.68 in narratives" vs "2.07 in academic journals" vs "0.93 in news magazines"). The distribution of coreference links is also regarded as one of the genre-specific characteristics because of these variances between genres that have been found.
- **Distribution of entity types**: The referents marked in the study are associated with the entity types of *action, concept, event, fact, object, organization, person, place, quality* and *time*. Observation of the distribution of referent types in the texts shows that texts belonging to the same genre have similar patterns. Differences between genres are shown by the distribution of the percentages of the most popular referent types. In academic journals, places and people are the first two most frequent entity categories referred to, in news magazines people and events, and in narratives, people and objects. Compared to other genres, academic journal texts display a more uniform distribution of referent kinds. This finding suggests that the distribution of referent types within each genre can be used to distinguish between them.

Amoia et al. [2012] explore the coreference variation in two genres from both spoken and written modes. The data contains interviews as instances of spoken mode and texts from popular science journals as instances of written mode. The motivation behind the study relies on the observed drop in the performance rates of automatic coreference resolution (ACR, henceforth) systems when they run on texts from multiple sources (e.g., web texts, conversations). For example, Luo et al. [2004] reported that their ACR system had a success rate of more than 80% for the MUC-6 dataset; but 7 years later, the highest ACR system performance score in CoNLL 2011 shared task is recorded as 57.7%. Although different evaluation metrics are used to calculate these scores, Amoia et al. [2012] argue that these values are still representative. They claim that inclusion of new genres (web texts and spoken texts) in the test portion of the CoNLL 2011 dataset is one of the primary causes of the observed performance decline. Because they can offer potential customised settings for the analyzed genres, Amoia et al. argue that contrastive studies exposing the genre and mode-based distinctions in coreferential patterns can benefit ACR research.

11 texts from both genres are examined in the study. The texts are annotated with the deterministic Stanford ACR system [Lee et al., 2011]. The significance of the differences in quantitative descriptive results presented in the paper is assured with Student's t-Test. The features used for the analysis and the relevant results are as follows:

- **Average token length of referring expressions**: Referring expressions are longer in written texts (3.42 vs 2.58).

- **Length of coreference chains**: Coreference chains in the spoken texts contain more expressions than the written texts (no quantitative result presented in the paper).
- **Distance between referring expressions** (in terms of sentences): The average distance is longer in written texts than in spoken texts (8.8 sentences vs 6.5 sentences).
- **Frequency distribution of referring expressions according to their grammatical roles**: In both modes, pronouns are more frequently employed in the subject position (86% in written texts and 97% in spoken texts). The grammatical roles of referring expressions of the NP type vary greatly. In written texts, referential NPs are more common in the subject position (41%) than in spoken texts, where they are most common in the object position (54%).
- **Frequency distribution of referring expressions according to their syntactic categories** (e.g. NP or pronoun): NPs are preferred over pronouns in written mode (63% NPs vs 29% pronouns in written texts vs 32% NPs vs 58% pronouns in spoken texts).
- **Frequency distribution of pronouns according to their morphological features** (in number and person): The most common pronoun type in both genres is the third person singular (45% in written texts and 38% in spoken texts). The frequency of first person singular pronouns in spoken texts is higher than in written texts (27% vs 6%), which is the most noticeable variation.

Amoia et al. [2012] argue that their findings indicate statistically significant differences between written and spoken texts and can be used to define criteria that are particular to a given genre and mode. For instance, the prevalence of first person singular pronouns in interview data illustrates how spoken texts are speaker-oriented. As a result, including speaker information as a parameter can benefit ACR systems for spoken mode.

The more recent studies on the variation of coreferential strategies explore the subject not only for different genres but also across languages and regional language varieties. **Lapshinova-Koltunski [2015]** conduct a corpus-based study investigating the intra- and inter-lingual variation of 3 linguistic phenomena: verb modality, coreference, and discourse relations. The motivation behind the study is to address the differences among languages in terms of highly used discourse phenomena for the improvement of NLP (e.g., machine translation) tasks. The dataset contains 406 texts from seven genres: political essays, fictional texts, instruction manuals, popular science articles, letters to shareholders, prepared political speeches, and tourism leaflets. The data is automatically tagged by computational tools. For the analysis of coreference, frequency-based statistics of demonstrative and personal pronouns and general nouns, e.g., *plan*, *case*, *fact*, are considered. In addition to descriptive statistics, an exploratory statistical technique, Correspondence Analysis [Greenacre, 2016], is used to identify differences between languages and genres as well as the potential relations between linguistic variables. In terms of the chosen features, Lapshinova-Koltunski expose both the cross-linguistic differences between German and English and the distinctive qualities of the genres. The findings suggest that coreference related features can be used to differentiate between genres. For instance, based on some of the linguistic characteristics taken into consideration, such as the high usage of referential general nouns, travel brochures, political essays, and popular scientific articles are grouped together. The frequent use of pronouns distinguish fictional works from other genres.

Kunz and Lapshinova-Koltunski [2015] study coreference (along with conjunction and substitution) in English and German written and spoken registers[25]. Original and translated texts belonging to ten different registers are annotated with a semi-automated procedure. Written registers include popular science texts, instruction manuals, and corporate websites. The two spoken registers are academic speeches and interviews. Personal pronouns, demonstrative pronouns, modifiers (e.g., this, that, here) and comparatives (e.g., similar, such) are explored as coreferential devices in the study.

Kunz and Lapshinova-Koltunski [2015] use Correspondence Analysis to uncover the differences and similarities between registers in terms of the investigated features. They find that fiction texts can often be identified by their high personal pronoun frequency. The findings additionally indicate that distinction of spoken and written modes is possible in terms of the investigated phenomena. For instance, spoken texts are distinguished by their frequent use of demonstrative pronouns, whereas written texts may be identified by their distribution of comparative referential expressions.

Kunz et al. [2016] investigate the interaction of identity relations (coreference chains) with other types of anaphoric relations (lexical chains, in their terminology) such as type-of and part-of relations. The included registers are political essays, popular science articles, fictional excerpts and transcribed interviews. Clausal referents are annotated in the data together with nominal referents. ANOVA is used for statistical significance testing and Correspondence Analysis for exploring the associations between features and registers. The following linguistic features are explored in relation to the included registers:

- **Length of coreference chains**: Popular science articles have the longest average lexical chain length (approximately 6). The other three registers are closer to each other in this respect with an average *mean* of 3.5. As for the coreference chains, fictional texts are more distinguishable with an approximate average length of 4. The other registers are closer to each other with average lengths shorter than 3.

- **Number of chains**: Essays include the fewest lexical and coreference chains, whereas fiction texts have the most of these chains.

- **Distance between referring expressions** (number of tokens): The average distance between referring expressions in the same lexical chain is 400 tokens for fiction, while it is 200 tokens for the other three genres. For coreference chains, fiction texts and interviews have the longest distance (around 400 tokens).

- **Semantic relation types** (e.g., identity, hyponym, synonym, etc.): While other semantic relations are more significant in essays and popular scientific texts, identity relations are more prevalent in fiction and interviews.

Kunz et al. [2016] observe that register-based differences are more distinctive than language-based differences. The groupings that emerge from correspondence analysis indicate that mode can also be a distinguishing factor. For instance, fiction texts and interviews can be characterized by longer average distance between referring expressions than other registers. Kunz et al. [2016] associate this grouping with the mode (spoken vs written) because fiction texts often include spoken-like elements such as dialogue.

Neumann and Fest [2016] account for register variation across regional varieties of English, in terms of three cohesive devices which are pronoun frequency, frequency of conjunctions and lexical density. The texts under examination are extracted from the International Corpus of English [Nelson et al., 2002] and belong to six varieties of English.

[25] When referring to the different text varieties they investigate, authors use the term “register.” We use their vocabulary to report on their research. For all of the studies included in this review, we adhere to this strategy.

The data include both spoken (broadcast discussions and conversations) and written texts (academic, administrative writings, and timed exams that are non-printed writings produced by students). Neumann and Fest [2016] employ the grand *mean*, i.e., average *mean* of the frequency of an investigated linguistic feature in the entire dataset, as the reference value to compare register specific values. The quantitative analysis demonstrates that, across all language varieties, the relative frequency of pronouns is above the grand mean in two spoken registers, broadcast debates and conversations, and below the grand mean in the written registers. Pronouns are used more frequently in the conversations ($median \approx 13$) compared to broadcast discussions ($median \approx 9$). Timed exams utilize pronouns more frequently than the other writing forms. The findings indicate that pronoun frequency can be utilized to discriminate between spoken and written texts because it is higher in spoken registers across all of the language varieties examined.

The theoretical framework for the discussion of the results is based on the functional analysis of language. Neumann and Fest [2016] attribute the differences among the registers of the same mode (conversations vs broadcast discussions) to the situation-based characteristics of the texts, such as the social distance between the participants.

Schnedecker [2018] investigates the impact of genres on referential expressions in French. She makes an empirical study, comparing news briefs and incipits of fairy tales. She shows that considering only the pronoun and noun frequencies (i.e., paradigmatic approach) is not adequate for differentiating these two genres which have situational proximity (i.e., they are both written texts, readers are not known and not present at the time of production, and they both have the aim of story telling). She proposed an alternative approach which she calls the “configurational” approach, which basically relies on the coreference chain features rather than the individual features of referential expressions. The following features are exploited in Schnedecker’s analysis:

- **referential density** (ratio of number of RE/number of words in the text): The fairy tales have greater referential density than the news.
- **average number of reference chains per text**: 3 for the fairy tales and 2 for the news
- **length of reference chains**: 7 referential expressions for the fairy tales vs 3.42 for the news
- **stability coefficient** (total number of different noun phrases used as anaphors): The lexical variations are more numerous in the news.
- **composition of the referential chain**: central characters of the fairy tales are referred to frequently by personal pronouns. The characters in the news makes greater use of full NP.
- **number of categories of nouns** (general nouns, nouns related to professions etc.): In the news, the variety of nouns are larger.

Lapshinova-Koltunski and Kunz [2020] investigate the effects of language, mode, and register on coreference phenomena. Seven different registers from the GECCo corpus[Kunz et al., 2021] both in spoken mode (academic speeches, general interviews) and written mode (political essays, literature, technical manuals, popular science and texts from company websites) are examined in the study. In addition to nominal coreference, the analysis also takes into account antecedents that denote events and facts, such as sentential antecedents, as in the example below.

(2.14) [We work for prosperity and opportunity]$_i$ because they’re right. [It]$_i$’s the right thing to do. [Lapshinova-Koltunski and Kunz, 2020, p.55]

The GECCo corpus is annotated for various cohesive devices, including coreference. Lapshinova-Koltunski and Kunz [2020] extract 34 features from the data and use these features in a prediction task and a classification task over the existing variational dimensions on the data: language, mode and register. A wide range of features including the morpho-syntactic features of anaphors (e.g., personal pronoun or demonstrative pronoun), grammatical functions of antecedents (e.g., subject) and features of coreference chains (length and number of coreference chains, distance between chain members) are used in the experiments. The experiments indicate that depending on the variational dimension, different sets of features turn out to be informative or distinctive about the dimension. Observations on the register dimension indicate that some registers (such as academic speeches) diverge significantly from the others, indicating a need for domain adaptation for the task of automatic coreference resolution for this genre.

Lapshinova-Koltunski et al. [2020] investigate the coreference phenomena in translated texts. They work on transcribed TED talks and news texts in English and their German translations extracted from the ParCorFull corpus [Lapshinova-Koltunski et al., 2018], which is annotated for coreference chains. In addition to nominal coreference, verbal and clausal referring expressions are also annotated in the dataset. They extract features such as morphosyntactic types of all mentions, chain properties (number of chains, distance between chain members measured in sentences) from the data and use exploratory data analysis techniques (i.e., correspondence analysis [Greenacre, 2016] and hierarchical cluster analysis [Everitt et al., 2011]) to examine the data. Their results indicate that variation across registers is more pronounced than variation across languages. Several contrastive results concerning the register differences are also reported in the study. For instance, in news texts the distance between mentions in a chain is longer than in the spoken TED talks. The cluster analysis shows that there are two distinct groups of features for differentiating language and registers. These groupings, according to the authors, show that while a more finely-grained classification of features distinguishes the register or the mode, more general features show the distinctions between languages (originals and translations, in their case).

Concluding Remarks

We made a review of comparative corpus-based studies investigating the variation in the establishment of coreference. The motivation behind the studies vary; some aim to account for the choice of referential forms either to establish a descriptive framework [Biber, 1992, Swanson, 2003, Kunz et al., 2016, Kunz and Lapshinova-Koltunski, 2015], or theory-driven analysis [Fox, 1987, Toole, 1996] or to distinguish the possible indicators of genres with the motivation to improve automatic coreference resolution systems [Amoia et al., 2012, Lapshinova-Koltunski, 2015, Lapshinova-Koltunski and Kunz, 2020, Lapshinova-Koltunski et al., 2020].

Regardless of the specific research question explored, all of the studies follow a data-driven methodology. The quantitative representation of the data is created with descriptive statistics with respect to the linguistic features relevant to coreference. The most common approach is to compare different domains in terms of the average statistical values obtained. However, average statistics can be misleading in some cases. Swanson [2003] argues that the average statistics may not be informative for the genres if the individual texts belonging to same genre do not have uniform distribution in terms of the quantitative features investigated. Therefore, Swanson compares three texts for each genre and interprets the results accordingly. In another study, as a complementary metric to average statistics, Biber [1992] uses the maximum values of features in analysis and shows that the correlation between the maximum and the average values could be useful for the interpretation of quantitative results.

In addition to descriptive statistics, some of the studies use exploratory techniques such as correspondence analysis and exploratory factor analysis to observe and explain the inter-relationships of the variation domains with respect to the linguistic criteria examined. The correspondence analyses of Kunz et al. [2016], Kunz and Lapshinova-Koltunski [2015] propose that the distinction between spoken and written modes is possible with the investigation of coreference-related linguistic features. For instance, Kunz et al. [2016] observe that fiction texts and interviews can be grouped together with respect to the distance feature and this grouping can be considered as an indication of mode sensitivity for the distance feature.

The heart of the comparative studies is the comparison of the linguistic domains (e.g., genre or mode) with respect to linguistic features such as the frequently used metrics below:

- **Distance between referring expressions and their antecedents**: Different distance metrics are used in the literature. Biber et al. [1999] and Kunz et al. [2016] measure the distance in terms of number of tokens, Fox [1987] in terms of number of clauses, Amoia et al. [2012] in terms of sentences, and Biber [1992] in terms of number of interfering references. The findings of the presented studies indicate conflicting observations in terms of the distance feature. Fox [1987], Biber et al. [1999], Kunz et al. [2016] argue that average distance is longer in spoken texts than in written texts whereas Biber [1992] and Amoia et al. [2012] claim the opposite.

  One source of the contradictory results could be attributed to the use of different metrics. Amoia et al. [2012] show that the sentence lengths in written and spoken genres may differ (25 tokens in written texts vs 20 tokens in spoken texts), so the distance in terms of tokens may not always correspond to the distance in terms of clauses or sentences. Therefore, even textual distance metrics are not always comparable. However Fox [1987] and Amoia et al. [2012] measure the distance in terms of similar units (clauses and sentences, respectively). Even their results display contradictory findings.

- **Length of coreference chains**: Biber [1992] and Amoia et al. [2012] indicate that spoken genres have longer chains than written genres. Kunz et al. [2016] make the similar observation for fiction; fiction texts contain more expressions than the other genres examined in Kunz et al. [2016].

- **Frequency of referential forms according to syntactic category** (NP vs pronoun): Pronouns are used more frequently in spoken mode than in written mode. Biber [1992], Biber et al. [1999], Amoia et al. [2012], Neumann and Fest [2016]. Lapshinova-Koltunski [2015] makes a similar observation for the genre type of fiction, which contains more frequent use of pronouns than the other written genres.

The findings indicate that coreference phenomena can serve as distinguishing factors for spoken and written modes. An interesting observation emerges with the examination of fiction as a genre type. Systematic similarities of this written genre with spoken data can be considered as an indicator of a spoken-written continuum (i.e., spoken and written texts constitute a continuous spectrum rather than a discrete binary distinction).

3

Corpora

In this chapter, we introduce the data sources utilized in our comparative corpus work and computational experiments. We exploit three corpora in this study. Two of them, i.e., OntoNotes (**ont**) and Switchboard (**swbd**), are widely used linguistic resources distributed by the Linguistic Data Consortium.[1] Details on these corpora are given in Section 3.2.1 and 3.2.2. The third corpus consists of conversational interactions on Twitter. This Twitter dataset, which henceforth will be called TwiConv or **tw**, is compiled from scratch for the purposes of this study and annotated for coreference. Section 3.1 contains a description of the corpus compilation and annotation processes, as well as a quantitative description of the TwiConv corpus.

Twitter's Developer Policy[2] prohibits the publication of tweet contents. Therefore, most Twitter databases only share unique tweet IDs and annotations, not tweet texts. However, if the corpus in question is tokenized using a a relatively complicated procedure or includes manual corrections, stand-off annotation layers (i.e., annotations are stored separately from the annotated document text) may not match the text content in the compiled corpus. We thus present a distribution method for TwiConv in Section 3.1.7 to map the original tweet texts with our annotations. To our knowledge, TwiConv is the first tweet corpus annotated for nominal coreference.

General statistics on these three corpora and coreference annotations they contain is presented in Table 3.1.

3.1 Twitter Conversations

We compile a new corpus composed of Twitter texts for the purposes of this study. Tweets on Twitter can be posted in one of two ways: as a reply to an earlier tweet or as a new non-reply tweet. Because we are interested in discourse level properties, we work on a corpus of tweets in context rather than individual tweets. The tweet context, in our study, is the conversation structure to which a tweet belongs, i.e., if it is sent as a reply or if it receives replies. Conversations on Twitter are organized into "tree" structures, as illustrated in Figure 3.1[5]. A single *thread* (in our terminology) is a path from the root to a leaf node of a conversation tree. We preserve only one thread from each conversation tree in our corpus to avoid any overlaps in tweet sequences, and we describe the details for the compilation procedure in 3.1.2. This choice is based on the assumption that each thread in

[1] `https://catalog.ldc.upenn.edu`

[2] `developer.twitter.com/en/developer-terms/policy`

[3] We count the clauses by applying the criteria introduced in Section 4.5.3.2.

[4] Only the sentences containing clauses are considered.

[5] Taken from the Twitter's official blog: `https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/building-and-serving-conversations-on-twitter`

	tw	**swbd**	**ont**	**tc**	**bc**	**bn**	**nw**	**wb**
# of files	185	147	1625	46	17	947	597	18
# of documents	185	147	2040	142	274	947	597	80
# of tokens	48172	248222	903467	103587	147118	225657	355641	71464
# of sentences	3503	30700	55570	14162	10798	12147	14786	3677
# of clauses[3]	6719	41786	110680	18242	21719	27219	35428	8072
# of non-singleton coreference chains	1734	6863	25872	2461	4518	8042	9328	1523
# of annotated mentions	6352	23541	103625	15345	20235	28103	34115	5827
average document length (token)	233.5	1688.6	442.9	729.5	537.0	238.3	595.7	893.3
average document length (sentence)	18.9	208.9	27.2	99.7	39.4	12.8	24.8	46.0
average document length (clause)	36.3	284.26	54.3	128.5	79.3	28.7	59.3	101.0
average sentence length (token)	12.3	8.1	16.3	7.3	13.6	18.6	24.1	19.4
average sentence length (clause)[4]	1.9	1.9	2.1	1.7	2.1	2.3	2.4	2.2
average clause length (token)	6.4	5.9	8.2	5.4	6.7	8.3	10	8.8
# of parenthetical clauses	71	4149	2636	1373	613	274	319	57
# of discourse markers	144	17K	12K	8193	3300	342	63	134
# of utterances not counted as clauses	213	8750	3933	3266	334	262	12	59
# of discourse markers	144	17K	12K	8193	3300	342	63	134

Table 3.1: General statistics on the corpora (**tw**: Twitter Conversations, **swbd**: Switchboard, **ont**: OntoNotes, **tc**: telephone conversations from ont, **bc**: broadcast conversations from ont, **bn**: broadcast news from ont, **nw**: newswire from ont, **wb**: web blogs from ont; *documents* are separate sections, if any, in a single *file*.)

a conversation tree is a coherent discourse that is independent of the other threads in the tree. In order to assess this hypothesis empirically, we analyze 25 complete conversation tree structures in terms of cross-dependencies between different branches of the trees and report the results in Section 3.1.1.

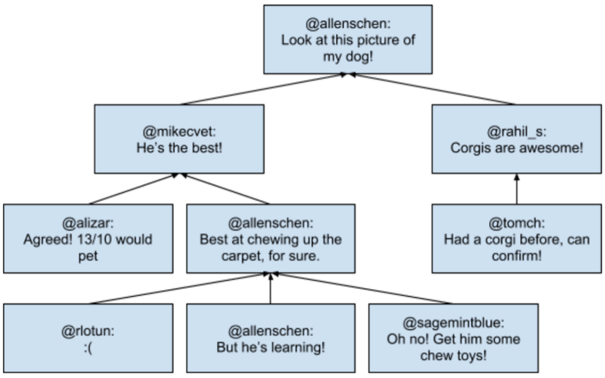


Figure 3.1: A conversation tree in Twitter

3.1.1 Cross-Dependencies and References in Twitter Conversation Trees

We construct a corpus of 25 twitter conversation trees to examine them in terms of cross-dependencies between different branches (i.e., threads). These 25 trees had a total of 4742 tweets with an average depth of 18 (ranging from 3 to 78) and average width of 21 tweets (ranging from 1 to 750). A tweet was on average 22 tokens (149 characters) long, or 20 tokens (111 characters) without the tagged usernames at the beginning of the tweet.

Conversation trees	25
Total number of tweets	4724
Average number of tweets	189
Total number of branches	2857
Average tree depth (tweets)	18
Average tree width (tweets)	21
Average tweet length (character)	149
Average tweet length (token)	22

Table 3.2: Descriptive statistics of the corpus compiled for dependency analysis

We distinguish two primary scenarios in which two threads are related. The first indicates that obtaining the meaning of one thread without the presence of another is not possible (i.e., cross-dependency). The second denotes a weaker relationship, in which there is at least one reference from one thread to another, but its meaning is still discernible (i.e., cross-reference). The following subsections provide illustrations of these scenarios.

3.1.1.1 Cross-Dependencies between Threads

Connectives There exist cross-dependencies resulted from discourse connectives, which were predominantly additive markers, more precisely either *and* or *also.* There were eight of those additive markers in all 25 conversations trees, six times *and* and two times *also*, used by four different users in four different conversations. Users utilize these markers when they replied directly to a tweet twice with both replies on the same level. In the instances we encountered:

- Except for one case, the chronologically second tweet contained the marker and additional comments or arguments to the first reply, and was posted in a time frame of one to seven minutes after the first (see in Figure 3.2).
- In one instance, a user replied with two same-level replies each to three different tweets in a time frame of seven minutes and connected all their tweets with additive markers (see in Figure 3.3).

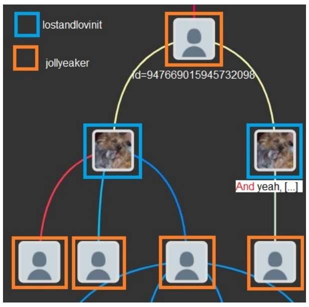


Figure 3.2: Cross-dependency Example 1

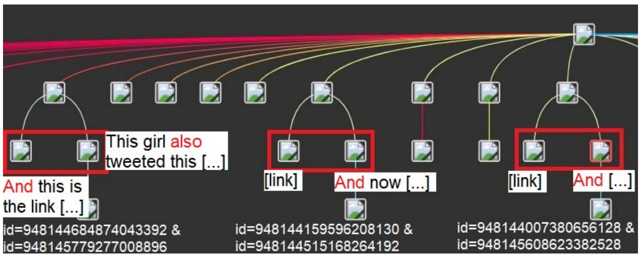


Figure 3.3: Cross-dependency Example 2

Enumerations Similar to using additive markers, users can intentionally post replies that would otherwise exceed the maximum character length as enumerated tweets on the same level, e.g., using "1/2" and "2/2" to indicate that the message is split up in multiple tweets. This happened twice in one conversation tree and the posters sent their responses using two or three tweets.

- Enumerating can affect the syntax because the tweets break apart sentences, e.g., "[...] You didn't like the 1/2" and "2/3 the scholarly article on police figures etc. [...]".
- Enumerating also affects coherence, e.g., "I gave you that Twitter thread to read 1/2" and "2/2 because I hoped it might make you walk a mile in someone elses shoes.".

3.1.1.2 Non-Dependent Cross-References between Threads

Several replies on the same level by one user can occur without necessarily creating cross-dependencies. The threads containing these replies are coherent on their own despite being often topic-wise related, e.g., when a user used three same-level tweets to provide the same evidence (see Figure 3.4).

In other occasions, references to earlier threads are used: two users in different trees refer to earlier conversation topics with "[...] someone in an earlier tweet said [...]" and "Refer back to what I said in one of my earlier threads".

3.1.1.3 Summary

We find 17 cross-references and dependencies in total. All of the cross-dependencies as well as some of the non-dependent crossing can be seen as a direct result of users deliberately choosing to reply more than once on the same conversation tree level. The same-level replies are especially obvious in enumerated tweets on the same level. Of all crossings and

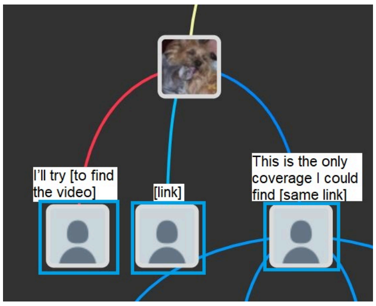

I'll try [to find the video]
[link]
This is the only coverage I could find [same link]

Figure 3.4: Cross-reference between threads

dependencies, 13 (76.5%) directly involved same-level replies which are typically expected to be on descending levels. The remaining instances are four non-dependent crossings which are a result of conversations by a small group of users that developed over several threads.

When compared to the size of the corpus (i.e., 4742 tweets), the number of dependencies and references (17 in all) is fairly small. As a result, we see this study as a confirmation of our hypothesis concerning the coherent structure of threads, even if the remainder of the conversation tree is removed from the data.

3.1.2 Corpus Compilation

After demonstrating that there are only few connections between various threads in a Twitter conversation tree, we construct a corpus of threads extracted from Twitter conversations. We gather new data from the Twitter stream for this purpose. We use *twarc*[6] to collect English-language tweets from the Twitter stream on several (non-adjacent) days in December, 2017. We do not filter for hashtags or topics in any way, since that is not a concern for this corpus. Conversations are gathered by recursively obtaining parent tweets, whose IDs were derived from the `in_reply_to_id` field of the tweet objects returned by Twitter API. We then use a script from Scheffler [2017], which constructs the conversational full tree structure for any tweet that generated replies. For the purposes of this study, we are not interested in alternative replies and other aspects of the tree structure; so we kept only one of the longest threads (path) from each tree and discarded everything else. A sample thread structure including a coreference chain annotation is illustrated in Figure 3.5.

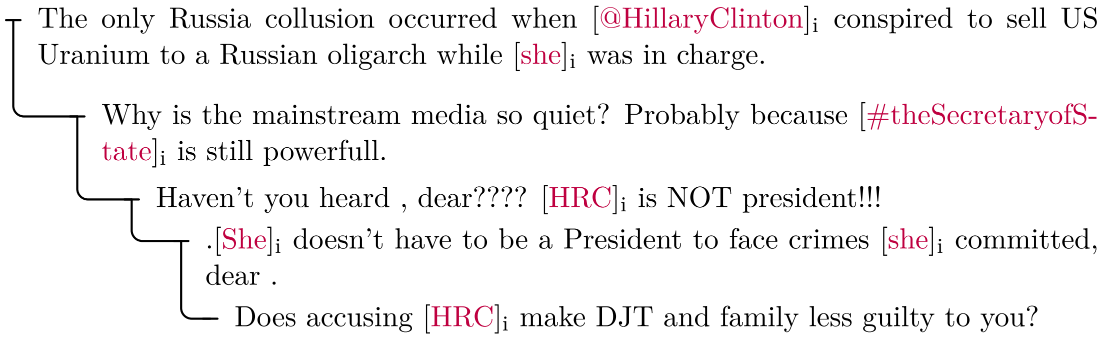
The resulting corpus consists of 1756 tweets arranged in 185 threads, and the average length of a tweet is 153 characters. The length of threads in terms of tweets varies between 3 and 78, with the average being 10 and median being 7. Quantitative summary of the corpus is shown in Table 3.3.

3.1.3 Tokenization

It is well known that tokenization is a crucial preparatory step for doing any kind of NLP. We experiment with two different tokenizers: the Stanford *PTBTokenizer* [Manning et al., 2014] and *Twokenizer* [Gimpel et al., 2011]. It turns out that these systems have different strengths in handling the variety of challenges. For instance:

- PTBTokenizer handles the apostrophes (e.g., possessive markers and contracted verb forms) whereas Twokenizer does not (rows 1-4 in Table 3.4 represent examples with an apostrophe and how PTBTokenizer and Twokenizer deal with them).

[6] `https://github.com/DocNow/twarc`



The only Russia collusion occurred when [@HillaryClinton]$_i$ conspired to sell US Uranium to a Russian oligarch while [she]$_i$ was in charge.

Why is the mainstream media so quiet? Probably because [#theSecretaryofState]$_i$ is still powerfull.

Haven't you heard , dear???? [HRC]$_i$ is NOT president!!!

.[She]$_i$ doesn't have to be a President to face crimes [she]$_i$ committed, dear .

Does accusing [HRC]$_i$ make DJT and family less guilty to you?

Figure 3.5: A conversation thread from the TwiConv corpus (strings in brackets compose of a coreference chain denoting the real world entity "Hillary Clinton")

# of threads	185
# of tweets	1756
# of tokens	48172
# of sentences	3503
# of clauses	6719
average thread length (token)	260.4
average sentence length (token)	13.6

Table 3.3: General statistics on the TwiConv corpus

- Twokenizer is stronger in recognizing the punctuation symbols such as sentence-final full stops, exclamation marks and also social media symbols (e.g., emoticons) even if they are not surrounded by white space. Some of these cases are illustrated in rows 5-7 in Table 3.4.

	String	**Twokenizer**	**PTBTokenizer**	**TwiConv Pipeline**
1	aren't	aren't (1)[7]	are, n't (2)	are, n't (2)
2	you've	you've (1)	you, 've (2)	you, 've (2)
3	London's	London's (1)	London, 's (2)	London, 's (2)
4	d'Orsay	d'Orsay (1)	d'Orsay (1)	d'Orsay (1)
5	here:)Because	here, :), Because (3)	here:)Because (1)	here, :), Because (3)
6	here.Because	here, ., Because (3)	here.Because (1)	here, ., Because (3)
7	U.S.	U.S. (1)	U.S. (1)	U.S. (1)
8	..	.. (1)	., . (2)	., . (2*)

Table 3.4: Tokenization outputs

We thus decided to implement a tokenization pipeline where the output of the Twokenizer is given as input to the PTBTokenizer. The outcome of this pipeline process is compatible with Penn Treebank conventions[8] and, therefore, with the other corpora following the same conventions, such as OntoNotes Weischedel et al. [2013] and Switchboard Calhoun et al. [2010]. One disadvantage of this method could be that it duplicates over-

[8] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

tokenization errors. For instance, as seen in row 8 of Table 3.4, the double dots (..) are correctly recognized as one symbol by the Twokenizer but then incorrectly divided into two tokens by the PTBTokenizer in the pipeline. Another example is the tokenization of the URL strings. Twokenizer considers string forms like *ftp://xxx.yyy* as URLs and treats them as a single token. PTBTokenizer, on the other hand, does not identify them as URLs and splits them up into multiple tokens. We discover that the quantity of tokens grew by 4% in the second step of the pipeline, where only 5% of newly issued tokens are erroneous over-generated tokens. Therefore, we do not consider over-tokenization as a threat to token-based compatibility with other corpora. In addition, over-tokenization is preferred to insufficient generation of token boundaries, because the annotation tool we used (i.e., MMAX) can handle markables with multiple tokes, but it does not allow for selecting a substring of a token as a markable. We also manually correct a few tokens in the output of tokenization pipeline.

3.1.4 Sentence Segmentation

We follow a semi-automated segmentation procedure to split the tokenized tweets into sentences. We first segment the text using SoMaJo, which is a sentence splitter for English and German web and social media texts [Proisl and Uhrig, 2016]. SoMaJo deals well with common Twitter tokens such as links, hashtags, and abbreviations but fails when sentences in the same tweet start with lowercase letter or hashtag, or when a user does not use any punctuation. Therefore, we manually correct the sentence boundaries detected by SoMaJo. Hashtags and links at the end of a sentence are considered as part of that sentence for the sake of consistency. We specified 3503 sentences in the TwiConv corpus.

3.1.5 Coreference Annotations

The TwiConv corpus has been annotated for identity coreference by undergraduate linguistics students co-operating with the project team, and the whole process has been supervised and coordinated by the author of this dissertation. We use the MMAX2 coreference annotation tool [Müller and Strube, 2006] for annotating the data. The tool is customized according to our scheme. We describe the annotation process and how we adapt MMAX to meet our requirements in Appendix C. The rest of this section covers the annotation principles and data quality assurance procedures.

Annotation Principles In our scheme, *markables* are phrases with nominal or pronominal heads. All nominal expressions, such as names, definite/indefinite noun phrases, pronouns, and temporal expressions are annotated for coreference. Non-referential pronouns (ex. 3.1), predicative constructions (ex. 3.2), and appositions (ex. 3.3) are also annotated and can be distinguished by the attribute values assigned to them. Elements of the web language such as usernames (ex. 3.4) and hashtags (ex. 3.5) are considered as markables as well. Links and emojis are treated according to their grammatical roles. We annotated all the chains including singletons in this corpus. Details of the annotation scheme and demonstrative examples are given in Appendix C.

Chains can contain several markables from the same tweet (intra-tweet) or from different replies (inter-tweet), which can lead to 1st, 2nd and 3rd pronouns referring to the same entity within one thread as in Example 3.6. We do not allow discontinuous markables, therefore split antecedents and their coreferring mentions are annotated as separate markables (Example 3.7) unless they occur as compound phrases (Example 3.8)[9].

(3.1) [It]'s raining today.

[9] The full guideline with examples is shared together with the corpus.

(3.2) [This]$_i$ is [a bank]$_i$, but [it]$_i$ is not very well-known.

(3.3) I called [Till]$_i$, [my friend]$_i$, to invite [him]$_i$ to join us.

(3.4) [@BarackObama]$_i$ should change [his]$_i$ policy.

(3.5) Yes! [She]$_i$ is my favorite. [#Oprah]$_i$

(3.6) Thanks to [you]$_i$, [I]$_j$ can now understand the whole conversation.
[You]$_j$ are welcome.

(3.7) [I]$_i$ met [him]$_j$ at [our]$_k$ favourite café.

(3.8) [The baby and I]$_i$ are listening to [our]$_i$ favourite music.

We define a comprehensive attribute set to be assigned to the annotated chains and mentions. For instance, all chains should be assigned a representative mention (i.e., the most descriptive mention in the chain), a semantic class (i.e., the semantic type of the entity) and a genericity value (i.e., whether the referred entity is specific or generic). Mentions are assigned its form (e.g., pronoun or a proper name) and grammatical role. For more details on the annotation scheme and for illustrative examples, see the coreference annotation guideline provided in Appendix C.

Quality Assurance We apply the following procedures to assess and improve the quality of manual annotations.

1. **Automated Checks** We validate the consistency of the annotations by applying a number of automated procedures checking whether the constraints specified in the guideline in Appendix C are applied uniformly. These automated procedures are applied for the consistency checks below. Instances of incorrect annotations detected by the automatic checks are manually fixed.

    - whether the nominal form (e.g., named entity, indefinite NP or pronoun) is assigned to all mentions
    - whether the pronoun type (i.e., anaphora, cataphora, exophora) is assigned to all pronouns
    - whether syntactic features are assigned to all mentions
    - whether semantic features are assigned to all chains
    - whether one mention in every chain is specified as the representative mention

2. **Inter-Annotator Agreement** We assess the inter-annotator agreement to evaluate the reliability of our annotation process. Only the coreference chains including 3rd person pronouns are annotated in the first version TwiConv corpus. We conduct an inter-annotator agreement evaluation on this initial version of the corpus. We then extend the guideline (GL) and annotate all coreference chains in the second version of the corpus. The changes in the extended GL concern only the attributes which are not addressed in the IAA study. We are thus confident that this agreement study can also assess our final scheme in terms of mention detection and chain linking.

    The agreement study is conducted on a sub-corpus composed of 12 randomly selected threads containing 9 tweets on average. This sub-corpus is annotated by 2 undergraduate linguistics students.

Artstein and Poesio [2008] evaluate different metrics for a variety of linguistic agreement tasks, including coreference. They argue that indications of percentage agreement are not fully reliable due to the impact of chance factor. The chance factor in percentage agreement is sensitive to two domain specific factors, which are the number of categories and the distribution of annotated items among categories. Thus, the results of percentage agreement cannot easily be used to compare different studies, according to Artstein and Poesio [2008].

The other common metric, kappa coefficient (κ) [Carletta, 1996], eliminates the chance agreement but computes the agreement by relying on exact matching of the annotated items. Artstein and Poesio [2008] show that for tasks such as coreference resolution, where reliability is determined by measuring agreement on sets (coreference chains) instead of discrete categories, disagreement should be evaluated gradually (i.e., partial agreement) instead of binary **match/do-not-match** decisions. Artstein and Poesio [2008] propose the use of Krippendorff's α [Krippendorff, 1980] for coreference agreement studies to overcome the limitations of percentage and κ metrics.

We compute the agreement of annotators for mention detection and chain linking. We follow the agreement evaluation method outlined and put into practice by [Kopec and Ogrodniczuk, 2014] and [Potau, 2010]. This method employs Krippendorff's α with weighted similarity comparison computed using MASI (Measuring Agreement on Set-valued Items) distance [Passonneau, 2006] between coreference chains. In order to construct the coincidence matrix for calculating the α value, as proposed by Passonneau [2004], we employ annotated entities as labels for the emerging categories. The similarity between the labels is computed by using the MASI distance. The resulting Krippendorff's α value is obtained as 0.872 ($\alpha \geq .800$) in this setting, indicating the validity of our data annotations for research purposes.

The most common annotator disagreements or errors result from different selection of mentions (missing or spurious markables), missing chains if they only contained very few mentions or the splitting of one chain into two, as well as occasional differences in markable span boundaries.

3. **Review of Annotations** We review a sample sub-corpus even though the agreement statistics indicate that the annotation process and the guideline provide a trustworthy foundation for coreference annotations.

   Annotations of first 27 threads (15% of all threads in the corpus) are reviewed by a graduate linguistics student, who was not involved in the annotation process. In this assessment, 33 problematic cases are identified, including 2 chains that were mistakenly merged, 1 chain that was overlooked, and 2 chains that had missing mentions. Most of the other problematic cases are the result of incorrect selection of mention spans or assignment of wrong attributes. Incorrect annotations affect approximately 50 mentions in total. There exist 2935 mentions annotated in this sub-corpus, including the singleton entities. As a result, the detected problems affect only 2% of all mentions in this sub-corpus. Therefore, we did not see the need to include the complete corpus in the review process.

3.1.6 Quantitative Description

We present descriptive statistics for the TwiConv corpus in Tables 3.1 and 3.3. Tables 3.5 and 3.6 provide statistics for the coreference annotations, which also contain singleton chains in the corpus. As the numbers indicate in Table 3.5, 61% of the coreference chains contain coreference links across tweets. Therefore, it is crucial for coreference annotations to take into account conversation context in Twitter messages. Twitter posts contain elements of web language such as usernames and hashtags. In this corpus, there are mentions including a username and 94 mentions including a hashtag. Links and emojis can also be included in mention spans if they are used as NPs.

# of chains:	7035
# of non-singleton (ns) chains:	1734
# of singleton chains:	5301
# of intra-tweet coref chains (ns):	674
# of inter-tweet coref chains (ns):	1060

Table 3.5: Descriptive statistics for the coreference chains

# of mentions:	12548
# of non-pronominal mentions:	7696
# of pronominal mentions:	4799
# of 1st Person Pronouns:	1215
# of 2nd Person Pronouns:	969
# of 3rd Person Pronouns:	1641
# of username mentions:	124
# of mentions including hashtag:	94
Average mention length (in tokens):	1.94

Table 3.6: Descriptive statistics for the mentions

In our annotations, non-pronominal and pronominal expressions constitute 62% and 38% of the annotations, respectively. As shown in Figure 3.6, definite noun phrases account for 60% of non-pronominal noun phrases, whereas other types of expressions (i.e., indefinite noun phrases and named entities) are distributed more evenly. Pronominal expressions, as illustrated in Figure 3.7, are mostly anaphoric expressions. Only 1% of the pronominal expressions are labeled as “cataphora”. Non-referential pronouns constitute 3.8% of all pronominal expressions. Figure 3.8 shows the frequency distribution for all pronominal expressions defined in our scheme. As shown in the figure, personal pronouns account for 70% of pronominal expressions.

The majority (81%) of the entities are marked as “specific” while the remaining 19% are categorized as “generic”. Human references account for 28% of the entity categories identified, followed by “abstract” entities at 10%. In terms of their semantic class, 44% of the entities are labeled as “other”, indicating that they do not fit into the categories we specified in our scheme (for more information, see Figure 3.9).

3.1.7 Distribution of Corpus

Due to Twitter’s Developer Policy[10], we have to refer to tweets via their IDs, through which the message text as well as other tweet-related information can be downloaded using the Twitter API. Below, we describe the procedure to compile the corpus from the tweet IDs and to align with the stand-off annotations we distribute.

3.1.7.1 Corpus Format

Annotations are stored in a CoNLL format[11] (i.e., tab-separated fields) with 17 columns in total, one file per Twitter thread. The content of each column is described in Table 3.7 and an example is presented in Appendix E. The Part-of-Speech tags and parses in column 4 and 5 are automatically created with the Stanford Parser Manning et al. [2014] with no manual correction. Empty lines indicate sentence breaks.

It is possible that different mentions start at the same token, e.g., “My Twitter username” marks both the beginning of the pronoun mention “My” as well as the full definite noun mention “My Twitter username”. In this case, we used pipe symbols (“|”) to separate the annotations for different mentions. The order of the annotations separated by the

[10] developer.twitter.com/en/developer-terms/policy

[11] https://universaldependencies.org/format.html

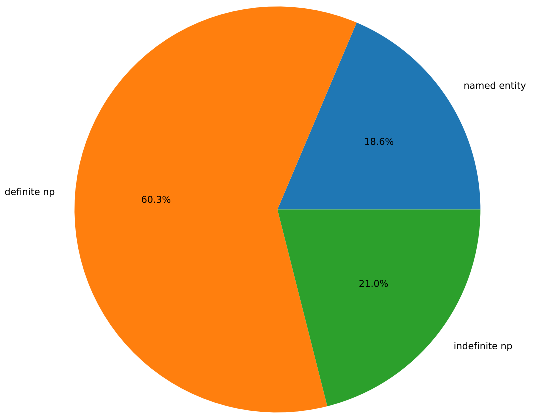


Figure 3.6: Non-pronominal expressions

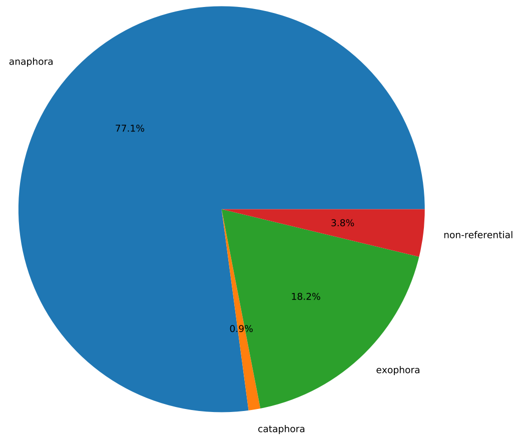


Figure 3.7: Pronominal types

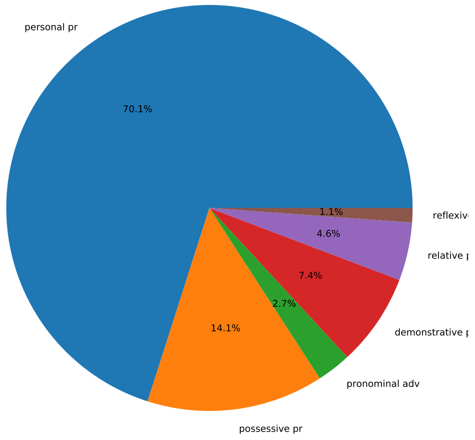


Figure 3.8: Pronominal expressions

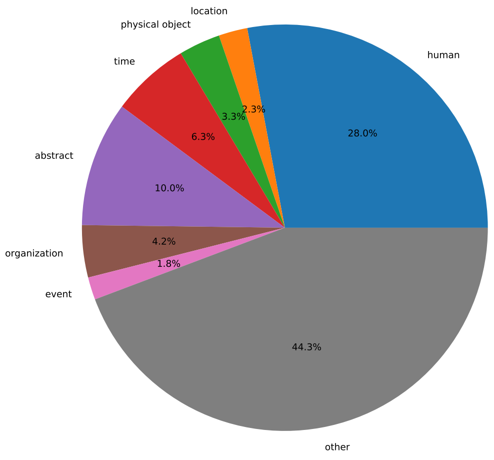


Figure 3.9: Entity types

pipe symbol remains the same for the entire line, meaning that the order of annotations in pipe-separated columns is always the same.

Further, some annotations such as NP form and grammatical role have sub-categories, which we express by slashes ("/"): e.g., *ppers/anaphora* marks a personal pronoun that functions as an anaphoric expression. Similarly, the grammatical role *other* can be either appositive, vocative or other (e.g., *other/vocative*), but those sub-categories were only

Column	Content	Column	Content
0	Thread ID	9	NP form/reference type
1	Thread No	10	Coreference ID
2	Token No in sentence	11	Clause boundary
3	Token	12	Shortest NP boundary
4	POS tag	13	Longest NP boundary
5	Parse info	14	Grammatical role
6	Speaker/User handle	15	Genericity
7	Representative mentions	16	[Tweet No in thread]_[Sentence No in tweet]_[Token No in sentence]
8	Semantic class		

Table 3.7: Column content in CoNLL format corpus

assigned to the *other* type, not to subjects, prepositional phrases etc.

We use the automatically created parses to detect the clause and NP boundaries (both for the shortest and the longest NP spans[12]) in tweets. We then manually correct these boundaries. The boundary start and end tokens are also included in the data files as specified in columns 11-13 in Table 3.7). The last column in the data files represents the relative order of tokens in the texts.

3.1.7.2 Sharing Method

In order to share the data, we use a method similar to the distribution of the CoNLL-2012 Shared Task Data [Pradhan et al., 2012] and provide skeleton files which include all annotations, but no tokens from the Twitter messages and no usernames (instead, they are replaced by underscore characters). For each token, the ID of the tweet from which the token originates is indicated at the end of the corresponding line. As we have tokenized the data, we also provide reference files to recreate our tokenization steps. To create those *diff* files, we compare files with the whitespace tokenized tweets (see Figure 3.10 for a sample representation), with one token per line, to ones with the tweets with our final tokenization, one token per line as well, with the Linux command *diff.* We share only those tokens in the *diff* files that were affected by the tokenization method or other forms of modification such as encoding differences for emoticons. After downloading the available tweets, they have to be transformed into the above described format (whitespace tokenized, one token per line, one file per tweet). We provide an assembly script that will use these tweet files, the skeleton files and *diff* files to create the complete CoNLL files with all annotations and tokens.[13] The script itself contains no information about the content of the annotations and can be re-used for any other tweets, given that the *diff* and skeleton files (following the CoNLL-style format described in Table 3.7) have been generated correctly. For unavailable tweets, the tokens will remain anonymized (meaning the underscore character remains).

[12] For details on the short and the long NPs, see Section 4.4.2

[13] Scripts and data to reproduce the corpus can be found at `https://github.com/berfingit/TwiConv`

```
This
is
just
a
test.
Hi
Twitter!
```

Figure 3.10: Example Tweet, whitespace tokenized

```
5,6c5
< test
< .
---
> test.
8,9c7
< Twitter
< !
---
> Twitter!
```

Figure 3.11: *diff* file example

3.2 Third-Party Datasets

3.2.1 OntoNotes

3.2.1.1 Corpus Description

The OntoNotes corpus [Weischedel et al., 2013] is composed of multi-language data (English, Arabic and Chinese) from a range of different sources and offers gold annotations at different linguistic layers such as part of speech tags, syntactic constituent parses and coreference chains. We conducted a variety of quantitative analyses on the coreference annotations of the English part of the OntoNotes corpus. The English portion of the OntoNotes corpus contains translations from Arabic and Chinese, as well as texts originally produced in English. For the purposes of this study, we only considered the original English data, in order to avoid the effects from potential translation divergences. The resulting data portion consists of both spoken and written language. Spoken data includes telephone conversations (tc), broadcast conversations (bc), and broadcast news (bn), whereas written data contains newswire texts (nw) and web blogs (wb). More specifically, the sub-sections of OntoNotes corpus are:

telephone conversations (tc): transcripts of informal conversations from the CallHome corpus

broadcast conversations (bc): transcripts of conversations in TV talk shows from CNN and MSNBC

broadcast news (bn): transcripts of broadcast news from ABC, CNN, NBC, MNB, Public Radio International and Voice of America (i.e., mostly edited language)

newswire (nw): texts from the Wall Street Journal

web blogs (wb): texts from web blogs and news groups

We use the CoNLL-formatted OntoNotes data Pradhan et al. [2013] and process it by using the open source library AllenNLP Gardner et al. [2017]. The data is organized into CoNLL 2011/2012-formatted *files*. In OntoNotes, long texts belonging to *bc*, *tc* and *wb* genres have been split into smaller parts (i.e., *documents* in CoNLL terminology) during the annotation process [Pradhan et al., 2012]. Therefore, there exist some *files* in the data which contain more than one *document*. The documents are annotated independently; there is no cross-document annotation in OntoNotes.

As shown in Table 3.1, the OntoNotes corpus contains 903K *tokens* distributed across 2040 *documents*. The size of the corpus and its subsections in terms of sentences, clauses[14] and tokens is also presented in the table.

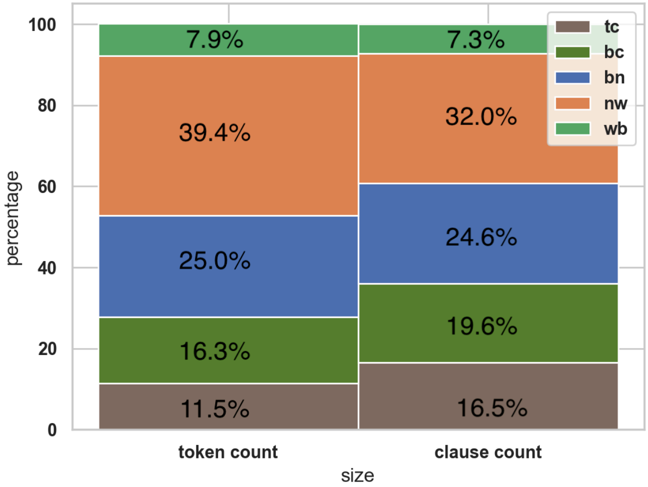


Figure 3.12: OntoNotes data size distribution

Measuring the data size in terms of different units results in different proportions. A visual representation of the data size distribution in terms of tokens and clauses is shown in Figure 3.12. The average length of the clauses differs according to the genre as shown in Table 3.1. Therefore, the data sizes also vary considerably according to the applied measure. For instance, in the token-based division of data in Figure 3.12, *tc* covers 11.5% and *nw* covers 39.4% of the data, whereas the proportion of the *tc* genre becomes 16.5% and the size of the *nw* decreases to 32% of the whole data in the clause-based comparison of genre sizes.

3.2.1.2 Tokenization

In the OntoNotes corpus, the texts are tokenized by following the Penn TreeBank tokenization scheme[15] [Weischedel et al., 2013]. The main principles of this scheme are as follows:

- Words and punctuation marks are considered as separate tokens.
- Verb contractions and genitive morphemes are separated from the root (e.g., children's → children,'s; I'm → I,'m).

Only the existing surface forms are considered as tokens in OntoNotes. For instance, silent moments and elliptical constructions are not marked as tokens in the transcribed OntoNotes data.

3.2.1.3 Linguistic Annotations

Texts in OntoNotes corpus have morphological (part-of-speech tags and lemmas), syntactic (constituents), semantic (argument structures, named entities, and WordNet senses) and discourse (coreference) level annotations. We are interested in the part-of-speech (PoS) tags, constituency parse trees and coreference annotations for the purposes of this study.

[14] We counted the clauses by applying the criteria introduced in Section 4.5.3.2

[15] `ftp://ftp.cis.upenn.edu/pub/treebank/public_html/tokenization.html`

PoS and constituency annotations follow the Penn Treebank2 conventions [Weischedel et al., 2013, Taylor et al., 2003] in OntoNotes.

Names, nominal mentions and pronouns are considered as markables in OntoNotes coreference scheme [BBN Technologies, 2007]. If a verb is used to refer as a noun phrase, such as in the example 3.9, it is also regarded as a mention.

(3.9) Sales of passenger cars [grew]$_i$ 22%. [The strong growth]$_i$ followed year-to-year increases.

In OntoNotes coreference annotations, identity (IDENT) and appositive (APPOS) relations are distinguished. For example, the appositive link between "Washington" and "the capital city" in "Washington, the capital city, is on the East coast" is marked by annotating "Washington" as the HEAD and the "the capital city" as the "ATTRIBUTE" of the APPOS relation. However, in the CONLL formatted OntoNotes that we process, the appositive links are not present. Instead, the HEAD and the ATTRIBUTE of the APPOS relations are merged. For instance, the string "Washington, the capital city" is considered as one single mention span instead of two mentions.

Singleton entities (i.e., entities referred only once), copula constructions, relative pronouns, most of the generic nouns and pre-modifiers are left out of the scope of the coreference annotation in OntoNotes.

Descriptive statistics on the coreference chains computed from the data are presented in Table 3.1. These chain and mention statistics give a good overview of the data that we are dealing with, but they should be interpreted carefully by considering the specifications in the OntoNotes annotation guidelines. Since the singleton chains are not annotated in OntoNotes, the statistics are not representative for describing the characteristics of genres in terms of referring expressions they contain. In addition to that, as stated in Pradhan et al. [2013], coreference annotation only covers document-level chains. As pointed out above, however, the texts belonging to *tc*, *bc* and *nw* genres are broken into multiple parts. Therefore, a more precise quantitative description of the genres in terms of referring expressions and chains could be computed only by providing the missing inter-document annotation of coreference, and by also marking the singletons; these steps would require substantial effort, though.

3.2.2 Switchboard

3.2.2.1 Corpus Description

Switchboard is a long standing corpus of conversational speech [Godfrey et al., 1992]. The original Switchboard corpus is composed of approximately 2400 spontaneous telephone conversations between unacquainted speakers of American English. The data is collected in an experimental setup where two strangers were given a topic from a pre-determined list and expected to have a conversation on the topic. Calhoun et al. [2010] bring together the existing annotations on Switchboard corpus and deliver a combined resource in NITE XML format[16].

The corpus contains 248K *tokens* arranged in 147 documents. The quantitative description of the data in terms of sentences, clauses and tokens is given in the Table 3.1.

3.2.2.2 Tokenization

The texts in Switchboard are tokenized by following the Penn TreeBank tokenization scheme. Silent moments, spots of grammatical ellipsis, repairs and false starts (i.e.,

[16] `https://groups.inf.ed.ac.uk/nxt/tutorials/tutorial1.shtml`

reparanda) are also included in the Switchboard transcriptions, which affect the number of tokens in the corpus. However, when we compute the token-based statistics in Table 3.1 for Switchboard, we only consider the existing surface strings in the text data (i.e., words and punctuation) and exclude the tokens indicating silences, elliptical constructions and *reparanda*.

3.2.2.3 Linguistic Annotations

The NXT format Switchboard has annotation layers for syntax, disfluency, speech acts, animacy, information status, prosody, focus/contrast, syllables/phones and coreference. The NXT Switchboard corpus includes 642 dialogs from the original Switchboard corpus, which are all annotated for PoS tags and syntax (i.e., constituency parses for each sentence). Among them, 147 of the dialogs are annotated for coreference. Since we are interested in the coreference annotations, we processed only these 147 files in our analysis.

In Switchboard, coreference annotation is done as part of the information status (i.e., the accessibility of entities in a discourse) annotation. As information status is a property of entities, only names, nominal phrases and pronouns are considered as markables. Information status is marked based on the hierarchy proposed by Prince [1992], where in a very simplistic description, the previously mentioned NPs are classified as *old*. For old entities, coreference links are marked between anaphor-antecedent pairs of the same entity. We construct complete chains from these pairs. Similar to OntoNotes, singleton entities, copula constructions, and relative pronouns are not annotated for coreference in Switchboard.

Coreference annotations include 23K mentions distributed to 6.8K coreference chains (see Table 3.1 for details).

4

Variation in Coreference: Corpus-based Evidence

4.1 Introduction

Research on strategies for producing referring expressions has often investigated the differences (if any) between spoken and written language, but as we cover in Section 2.3 and briefly summarize again in Section 4.2, findings are not always compatible. Sometimes, claims are simply contradictory, but the more important problem is that usually, the precise methods for computing the investigated measures are not being made transparent. Additionally, it is not always evident how different studies can be compared because the data they utilize might differ greatly. Our aim here is to gather empirical evidence that can contribute to clarifying the picture.

Our primary goal here is to shed light on coreference with respect to the spoken-written continuum, by undertaking a careful comparative corpus analysis and clearly describing our methods of measurement. We present a quantitative study on different genre sections of the OntoNotes corpus (**ont**), which is composed of spoken and written texts, and the Switchboard corpus (**swbd**), which is composed of spoken telephone conversations.

The secondary goal is to explore how the medium microblog, specifically Twitter, relates to the spoken-written spectrum for coreference patterns. In Chapter 6, we investigate this research question empirically through computational experiments on a state-of-the-art "standard" coreference resolution system. We show that the choice of genre and the mode (spoken vs written) in training data can make a bigger difference than the bare amount of data. In the current chapter, we apply corpus-based empirical methods to situate the coreference patterns found in Twitter conversations in the mode/genre spectrum. For this purpose, we include in this corpus study the comparison with Twitter texts that are included in the TwiConv (**tw**) corpus, thus achieving a wide range of production mediums.[1]

Although Switchboard and OntoNotes have previously been used for investigating coreference, to our knowledge they have not yet been systematically compared. Accordingly, an important part of our work is in harmonizing the data sets and the underlying annotation schemes, to enable a sensible analysis. Section 4.3 describes the corpus alignment processes we follow.

We will show genre-specific distributional patterns of nominal referring expressions in terms of **frequency** of syntactic categories (Section 4.5.1), **heaviness** of NP structures (Section 4.5.2), and relative **distance** between anaphors and antecedents in the text (Sec-

[1] The name of the files from OntoNotes and Switchboard that are used in this study, as well as the script to compute the TwiConv statistics can be found in the repository `https://github.com/berfingit/coreference-variation`.

tion 4.5.3). Most of our analyses lead to a common ranking and clustering of the genres based on the measures and results, as will be discussed in Section 4.6. Section 4.7 draws some concluding remarks.

4.2 Related Work

Various linguistic coreference phenomena have been compared by researchers in different domains such as across languages (e.g., Lapshinova-Koltunski [2015], Kunz and Lapshinova-Koltunski [2015], Engell [2016], Kunz et al. [2016]), regional language varieties (e.g., Neumann and Fest [2016]), production modes (spoken vs written) (e.g., Fox [1987], Biber [1992], Amoia et al. [2012]) and across genres in these domains. Among the features, frequency-based statistics and distance measurements are the most prominent. For distance between referring expressions and their antecedents (the closest previous mention of the same referent), different metrics have been used in the studies, which yield partly incompatible results:

- Biber et al. [1999] and Kunz et al. [2016] measure the distance in terms of number of tokens. The findings of these studies are similar to each other: The average distance is longer in spoken texts than in written texts.
- Fox [1987] measures the distance in terms of number of clauses, and argues that the average distance is longer in spoken texts than in written texts.
- Amoia et al. [2012] measure the distance in terms of sentences, and argue that the average distance is longer in written texts than in spoken texts.
- Biber [1992] measures the distance in terms of number of interfering references and concludes that the average distance is longer in written texts than in spoken texts.

These partly-incompatible findings indicate that textual distance metrics are not easily comparable; for instance, the distance in terms of tokens may not always correspond to the distance in terms of clauses or sentences. This is one of the aspects we will address in this chapter.

Other coreference phenomena that have been studied include the distribution of referring expressions in terms of their syntactic categories, i.e., pronouns vs noun phrases (NPs). Fox [1987] argues that referential NPs are generally more frequent in written texts than in spoken conversations (47% in written texts vs 22% in conversations), whereas Biber et al. [1999] and Amoia et al. [2012] observe different characteristics: On the one hand, they confirm Fox's finding that NPs are more frequently used than pronouns in the written mode than in the spoken mode. But in written texts, according to Amoia et al. [2012], NPs are more frequent than pronouns as well (63% NPs vs 29% pronouns), which is not in line with Fox [1987] who argues that pronouns are more frequent than NPs for both modes. Concerning the length of the NPs, they find that the average number of tokens of referring expressions in the written data is longer than the spoken data (3.42 tokens vs 2.58 tokens). The other commonly used quantitative metrics of coreference patterns in the literature are the number of referring expressions [Biber, 1992, Schnedecker, 2018], the number of referents [Biber, 1992, Kunz et al., 2016], and chain length [Biber, 1992, Amoia et al., 2012, Kunz et al., 2016, Schnedecker, 2018]. We do not examine these features because in OntoNotes, documents are artificially split in smaller parts, and because singletons are not annotated (see Section 3.2.1); and in Switchboard, there exist unannotated coreference chains in the data (see Section 4.3). Therefore, these metrics can create misleading results.

Although it is heavily used for coreference research, OntoNotes has, to our knowledge, not been extensively examined for quantitative comparison of reference features across genres. Among the exceptions is Hardmeier et al. [2018], who investigate how organizational entities are being referred to, and find a correlation between preferred reference type and genre (e.g., pronouns are more common in telephone conversations than newswire and broadcast news). Zeldes [2018] uses OntoNotes for predicting 'notional anaphora' (i.e., pronouns disagreeing with their antecedents' grammatical categories for notional reasons, e.g., "the government" and "they"), and find it to be more common in broadcast conversations than in newswire. Zeldes employs 20 linguistic features, and genre emerges as the third-most important one, indicating that differences between genres can have an impact on automatic classification.

As we report in Chapter 6, the state-of-the-art systems vary considerably on different domains. Most of the cross-domain experiments are done on the different datasets. There is little work that specifically addresses the performance difference of coreference systems across genres existing in the same dataset (i.e., tokenized and annotated with the exactly same schemes). We perform experiments with the Berkeley coreference resolution system Durrett and Klein [2013] on the data that we are dealing with and observed similar performance differences according to genre. After dividing the TwiConv dataset described into train and test sets, we train the Berkeley resolver on the train set and test the system performance for the different genre sections of the data. The F1 scores[2] according to genre are as follows: broadcast conversations 56%, broadcast news 63%, newswire 60%, telephone conversations 63%, and web texts 56%. Although the data size for each of the genres varies considerably, we still consider these numbers as possible indicators of variation in the coreference properties of different genres.

4.3 Corpus Homogeneity

Transcription The spoken texts can differ in terms of the transcription procedures applied. For instance, in Switchboard, *silence* moments and *traces* of grammatical ellipsis are inserted as separate tokens into the transcribed texts, whereas in Ontonotes, only the surface linguistic forms and punctuation are considered as tokens. In addition to this, unlike OntoNotes, repairs and false starts (i.e., reparanda) as illustrated in Example 4.1 are also included in Switchboard transcriptions. We do not take these additional tokens in Switchboard into consideration in this study. As a result, the sentence in 4.1 is formatted as in 4.2 in our analysis. Tokens marked by a *META* tag in OntoNotes, which are referring to the metadata of the texts, such as the "Reporter" of broadcast news, are not considered in this analysis either.

(4.1) $\text{I}_{\text{reparandum}}$ *SILENCE* I 'd like *TRACE* to see something like that .

The same sentence would be transcribed in OntoNotes without SILENCE, TRACE and reparandum tokens as below:

(4.2) I 'd like to see something like that .

Tokenization All the investigated corpora follow PTB tokenization conventions[3] with various adaptations based on the specific string types included in the texts. For instance, smileys (e.g., ":)", ":-("), emojis, hashtags (#TIMESUP) and links (https://t.co/Bgyj3U71HK) are considered as single tokens in Twitter texts, which would be handled in a different way

[2] The performance rates are calculated with the CONLL scorer as explained in `http://conll.cemantix.org/2011/faq.html`

[3] `ftp://ftp.cis.upenn.edu/pub/treebank/public_html/tokenization.html`

in the standard PTB tokenization scheme. The usernames of the conversation participants in **tw** thread structure, introduced by the @ sign, are automatically added to the content of the reply message in Twitter. Since these are not inserted to the post intentionally by the user, we consider such usernames ($\approx$ 5K in total) as part of the metadata of the tweet and do not count them as tokens in the text.

Linguistic Annotations OntoNotes and Switchboard have gold part-of-speech (PoS) and syntax (constituency parse trees) annotation layers compatible with Penn TreeBank conventions Taylor et al. [2003]. TwiConv does not include gold annotations for these layers. We thus use the Stanford parser Manning et al. [2014] to automatically create the PoS and syntax annotations for Twitter texts, which are also compatible with PTB conventions. However, the predicted parses are not reliably accurate for tweet texts[4], and therefore we manually correct the linguistic information (e.g., clause and NP boundaries) computed through these parses in our analysis. There is one case where the manual correction is considered as too costly in the computation of one metric related to syntactic complexity of NPs (Section 4.5.2). In that case, we do not apply a manual correction step. However, to increase the data quality, we exclude the NP structures which we consider as interpreted wrong by the automatic parser. Detailed information on the procedure can be found in the relevant section.

The other annotation layer of interest is the coreference annotations. All three corpora contain gold annotations for coreference, but with various differences in the definition of markables. For instance, in OntoNotes and Switchboard, singletons, copula constructions, headless relative clauses and appositions are not annotated [Pradhan et al., 2007, Calhoun et al., 2010]. Hence, for compatibility, we ignore these types of mentions in the **tw** corpus when comparing it with the other corpora.

As specified in the OntoNotes guidelines [BBN Technologies, 2007], verbs are annotated as mentions in the OntoNotes corpus when they refer to the same entity as a nominal mention. An example chain containing a verbal entity is "chain_meeting=[met, the meeting, the APEC meeting, it]". Since we want to focus on nominal coreference (which is also in line with the majority of work in coreference resolution), we exclude these non-nominal mentions in OntoNotes in our analyses.

Another difference in the coreference annotation schemes is that only markables with information status "old" are annotated for coreference in Switchboard [Calhoun et al., 2010]. However, not all candidates of referring expressions compatible with the markable definition are annotated for information status. This indicates that the annotated coreference chains do not cover the complete set of non-singleton entities in Switchboard. Therefore, we chose not to include the cumulative metrics of referring expressions (number of coreference chains, number of mentions) in our comparative analysis. Table 3.1 shows the summary statistics for annotations, but they are not fully comparable due to this design preference in Switchboard. Although the selection criteria for markables are not clearly described in the Switchboard documentation, we assume that the annotated chains are internally complete (i.e., all mentions for an annotated entity are marked), and therefore, can serve our purposes in terms of the chain-internal features we investigated (i.e., distance-based comparison in Section 4.5.3).

Remaining Incompatibilities Although we tried to align the datasets as much as possible, there still exist incompatible categories in the data. For instance, all datasets deal with "generic nominals" in different ways, which we demonstrate through a bare plural nominal *parents* in the examples below.[5]

[4] Analysis on the performance of automatic parsers on Twitter texts can be found in [Abbas, 2015].

[5] Examples are adapted from Pradhan et al. [2007]

(4.3) [Parents]$_i$ should be involved with [their]$_i$ children's education. [..] If [parents]$_k$ are dissatisfied with a school, [they]$_k$ should have the option of switching to another. (OntoNotes, two chains)

(4.4) Parents should be involved with their children's education. [..] If parents are dissatisfied with a school, they should have the option of switching to another. (Switchboard, no annotation)

(4.5) [Parents]$_i$ should be involved with [their]$_i$ children's education. [..] If [parents]$_i$ are dissatisfied with a school, [they]$_i$ should have the option of switching to another. (TwiConv, all in same chain)

4.4 Data Processing

The main resource for computing the quantitative metrics in this study is the linguistic annotation layers existing on the texts. As described in Chapter 3, the three corpora we used in this analysis (i.e., OntoNotes, Switchboard and TwiConv) are manually annotated for coreference. However in OntoNotes and Switchboard datasets, the singleton chains (i.e., entities referred by only 1 referential expression in the text) are not annotated. Therefore, evaluating only the properties of annotated referential expressions would be misleading because of excluded singleton chains. For the purposes of this study, we prefer to analyze the quantitative features of all nominal expressions in heaviness and frequency based comparisons, regardless of their referentiality. The drawback of this approach is including the non-referential noun phrases in the analysis, hence, of not delimiting the expressions relevant to coreference. In the TwiConv corpus, we detect that 3.5% of the annotated personal pronouns are non-referential (e.g., "**It** is raining" or "You can make **it**!"[6]). Given that this rate could serve as evidence for the prevalence of non-referential noun phrases in the corpora that we examine, we do not expect that including non-referential expressions in our analysis will have a significant impact.

For the analysis we first detect the pronouns and non-pronominal noun phrases. The detection procedures are described in the sections that follow.

4.4.1 Detecting Personal Pronouns

We capture the pronouns in OntoNotes and Switchboard through the PoS tags of tokens (i.e., pronouns are assigned the tag **PRP**). In Twitter, there is no gold PoS or parse information. Therefore, to capture the personal pronouns in the text and to differentiate their syntactic roles (e.g., accusative or possessive), we use the attributes in gold coreference annotations. As described in Twitter annotation guidelines in Appendix C, all the personal pronouns are assigned their syntactic roles in Twitter data. After the detection of a pronoun, we differentiate it in terms of the "person" feature (i.e., 1st, 2nd and 3rd person) by matching its string form with the list of pronouns we specified.

Personal pronoun (PRP) forms common to all corpora included in this analysis are shown below[7]:

1. 1st Person Pronouns: "i", "me", "mine", "myself", "we", "us", "ourselves", "'s", "ours"

2. 2nd Person Pronouns: "you", "yourself", "yourselves", "yours", "y'all"

[6] The pronoun is non-referential because it is part of an idiom

[7] The varieties are shown in lower-case form because case-insensitive string match is applied in the pronoun search.

3. 3rd Person Pronouns: “he”, “she”, “it”, “they”, “him”, “her”, “them”, “himself”, “herself”, “itself”, “themselves”, “his”, “hers”, “theirs”, “its”
4. 1st Person Possessive Pronouns: “my”, “our”
5. 2nd Person Possessive Pronouns: “your”
6. 3rd Person Possessive Pronouns: “his”, “her”, “their”, “its”

Along with these widespread variations, corpus-specific variants also exist. These may come from transcription preferences or the choices made by the speaker or writer in the source texts. The list below includes these forms. As a common pattern in Twitter texts, we observe contracted forms of nominative pronouns such as “im”, “ur”, “hes”, and “youve”. Since no apostrophe is used in these contractions, they are treated as single tokens and hence variations of personal pronoun forms.

1. Twitter:
   (a) **1st Person Pronouns:** “im”, “id”
   (b) **2nd Person Pronouns:** “u”, “ur”, “youre”, “ya”, “youu”, “youve”
   (c) **3rd Person Pronouns:** “hes”, “shes”, “em”
2. Switchboard:
   (a) **1st Person Pronouns:** “i-”, “w-”
   (b) **2nd Person Pronouns:** “you-”, “yo-”, “y-”
   (c) **3rd Person Pronouns:** “sh-”, “‘em”, “em”, “th-”, “the-”
3. OntoNotes:
   (a) **1st Person Pronouns:** “i-”, “w-”
   (b) **2nd Person Pronouns:** “yo-”, “y-”
   (c) **3rd Person Pronouns:** “sh-”, “‘em”, “em”, “th-”, “the-”

4.4.2 Detecting Noun Phrases

An NP span is any portion of text that has an NP structure. However, noun phrases can be embedded in each other, and hence the boundaries can differ according to the detection procedure applied. For instance, the string “an invasion of the privacy” can be recognized as one single *large span* NP, two *short span* NPs (“**an invasion**” and “**the privacy**”) or three NPs (“**an invasion of the privacy**”, “**an invasion**” and “**the privacy**”). We compute NP-based metrics both considering the large and the short NP spans.

As OntoNotes and Switchboard include gold annotations of constituency parse trees, we use these parses to extract the NP structures. The details of methods applied to find the NP spans are given in the sub-sections below.

Largest NP Spans The largest possible noun phrases in a text are considered as the NP boundaries (Figure 4.1.a). To detect the largest NP spans, we traverse the sentence parse trees in a top down **breadth-first** manner and extract the **first** encountered NP nodes in each branch. For instance, in example 4.6, “an invasion of the privacy” is considered as the largest NP span.

(4.6) I guess that’s [an invasion of the privacy]. (taken from **swbd**)

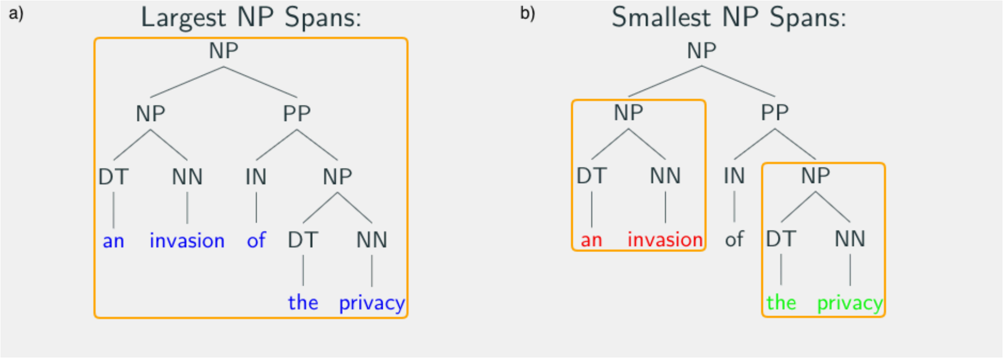


Figure 4.1: NP text spans

Shortest NP Spans The shortest possible noun phrases in a text are considered as the NP boundaries (Figure 4.1.b). To detect the shortest NP spans, we traverse the tree in a top down **depth-first** manner and extract the **last** encountered NP nodes in each branch. For instance, in example 4.7, "an invasion" and "the privacy" are considered as two shortest NP spans in the sentence.

(4.7) I guess that's [an invasion] of [the privacy]. (taken from **swbd**)

As TwiConv does not have gold PoS and syntax annotations, we apply a semi-automated procedure for identifying the NP boundaries in Twitter texts. After creating the constituency trees with the Stanford parser and detecting NP boundaries in the trees using the automatically created parses, we correct all the recognized NP spans manually. This step is necessary, because we observe that 75% of the short span NPs is identified correctly by the predicted parses, in contrast to only 40% of the large span NPs. The details of the procedure are described in Section 4.4.3.

4.4.3 Detection of NP Boundaries in Twitter

We apply the following procedures to the non-standard texts in Twitter posts.

- **Emojis and smileys** are replaced by "%emoji" string before running the automatic parser on the corpus. These strings are usually tagged as NP by the automatic parser, but we do not consider them as an NP unless they have a syntactic role of an NP in the sentence. For instance, in example 4.8, the emojis at the end are not considered as NPs, but in 4.9, the emoji has a syntactic role of an NP, therefore it is counted as an NP. There are 4 cases (among 553 emojis) in the **tw** corpus, where an emoji is considered an NP or part of an NP.

  (4.8) `i guess both are gonna suffer heavy duty` 🤒🤕

  (4.9) `[`🐍`] are fools...`

- **Links** (i.e., URLs) are usually tagged as NP by the automatic parser, but we consider them NPs only if they have a syntactic role of an NP in the sentence. For instance, in example 4.10, the link is not considered as an NP, but in 4.11, the URL has the grammatical role of an NP. There are 4 cases (among 340 links) in the **tw** corpus where a URL is considered as an NP.

  (4.10) As the president acts more erratically and lawlessly, GOP politicians are somehow growing * more * devoted to him. **https://t.co/Bgyj3U71HK**

(4.11) If crashing, please refer to this: [**https://t.co/NCvwPFGeaM**]

- **Usernames**, introduced by the "at" sign, **@**, are usually tagged as NP by the automatic parser. The automatically-inserted usernames at the beginning of reply tweets are not considered as NPs (example 4.12) in our analysis. However, we see usernames that are integrated in sentence syntax (4.13) or at the end of a tweet (4.14) purposefully added into the tweet text. We thus count them as NPs. This holds for 179 of the 5K usernames in the **tw** corpus. The others are automatically inserted trailing usernames that are not considered inside the NP boundaries.

(4.12) @brycetache @TeresaMac2009 @BarackObama Thank you Obama![8]

(4.13) [@JoeNBC] just said twice [the @washingtonpost] deleted his unnamed quotes from someone around Trump saying he exhibiting signs of dementia.

(4.14) Many websites do not load on safari. why? What is happening? [@AppleSupport]

- **Hashtags**, introduced by the hash symbol, #, are considered as NPs in the **tw** corpus. There are 205 instances of hashtags in **tw**, which we all recognize as NPs. However, handling strategy differs according to the syntactic function of the hashtag: We regard those with a syntactic role of an NP in sentence structure as separate NPs (e.g., #SecretaryofState in (4.15)). Hashtags that are not syntactically integrated and placed at the beginning or end of a tweet are also considered as NPs; but in case of more than 1 consecutive hashtag (also in (4.15)), they form a single NP. The content of a hashtag may contain several words (e.g., #SecretaryofState), but we do not do any segmentation. Hence, regardless of actual content, hashtags are always considered as a single token in all our computations presented below.

(4.15) The only Russia collusion occurred when @HillaryClinton conspired to seel US Uranium to a Russian oligarch while she was [#SecretaryofState]. [#RussiaCollusion #UraniumOne]

4.4.4 Additional Concerns: NPs in Parenthetical Clauses

Filler clauses such as "I mean", "you know" as in the examples 4.16 and 4.17 are frequent in spoken genres. These **parenthetical** clauses are discourse markers [Fox Tree and Schrock, 2002], and hence not as strong as other clauses in terms of adding new referential expressions into the discourse model. The parenthetical clauses are marked with **PRN** tag in gold constituency parses of OntoNotes and Switchboard. We compute additional statistics in terms of the described metrics in a setting where parentheticals, and hence the NPs they contain, are excluded. We use the **PRN** tag to detect the parenthetical clauses in all corpora, including Twitter for which automatically created parses served for this purpose with no manual correction for this particular task. In addition to the deictic "I" and "you" pronouns, parenthetical clauses can also contain other pronouns as given in examples 4.18 and 4.19 or full NPs as in 4.20. Descriptive statistics on the frequency of these structures can be found in Table 4.1.

(4.16) [I mean] I think legally he's not hopeless.(tc)

(4.17) Well [you know] that brings up the interesting subject too [you know]. (swbd)

(4.18) And I'm going to Vermont, [believe it or not], for graduate school.

[8] The initial usernames are inserted automatically by the Twitter interface as all these usernames are mentioned in the previous message replied to.

(4.19) We're confident that it protects our route structure, [he says], and our ability to grow and prosper.(nw)

(4.20) But because first-quarter demand is normally the weakest of the year, [several market participants say], OPEC production will have to decline to keep prices from eroding further.

4.5 Data Analysis

We use frequency-, heaviness- and distance-based metrics in the quantitative comparison of the (sub-)corpora surveyed in Table 3.1.

4.5.1 Frequency-Based Features

Description As frequency-based metrics for the genre/mode comparison, we compute the relative distribution of nominal expressions according to their syntactic categories (i.e., PRPs vs NPs) and distribution of PRPs according to the grammatical person feature (i.e., 1st, 2nd and 3rd person PRPs).

Results Frequencies of pronouns and non-pronominal noun phrases are given in Table 4.1 and distributions are shown in Figure 4.2 and 4.4 in terms of large span and short span NPs, respectively. Because possessive pronouns are not the head of noun phrases (NPs), but instead are a part of such phrases, they are not included in the computation of personal pronoun frequencies.

Table 4.1 shows that pronouns in parenthetical clauses are frequent in conversational spoken genres. In Figure 4.3 and 4.5, we present the syntactic category distributions of NPs and PRPs when parenthetical clauses are excluded. The Pearson's χ^2 with post-hoc pairwise Fischer test where the correction method for multiple comparison is set to "holm" indicates that the differences in the proportions of syntactic categories across genres (shown in Figure 4.2 and 4.4) are statistically significant ($p<0.05$) as shown in Table A.1 in Appendix A. When we exclude the nominal expressions in parenthetical clauses the differences except from "tw-bc" pair are still significant.

Figures from 4.2 to 4.5 demonstrate that the NP and PRP proportions are similar for spontaneous conversational spoken genres, and the gap between the proportions of NPs and pronouns are larger in written genres where NPs are more frequent than PRPs. The edited **nw** texts contain the highest proportion of NPs. Twitter (**tw**) texts lie in between the written and spoken genres[9] and closer to the **bc** in terms of the frequency distributions.

We also compare the relative frequency of NPs and 3rd person pronouns to get a more insightful view on the usage of anaphoric devices. The comparative graphics are shown in Figure 4.6 and 4.7. The statistical significance tests (results given in Table A.1 in Appendix A) indicate that all the genre pairs except from "bn-wb" have statistical significant differences. Usage of non-pronominal nominals is much more frequent than 3rd person pronouns in all genres. However, we observe higher usage of 3rd person pronouns in spoken genres than in written genres. Similar to the results demonstrated in Figures 4.2-4.5, **tw** lies in between the spoken and written genres also when only the 3rd person pronouns are considered in comparison.

In addition to the distribution of syntactic categories, we also compute the frequency distribution of personal pronouns according to their "person" type (i.e., 1st, 2nd, 3rd). Possessive pronouns are also considered in this frequency-based comparison of pronoun types. The representative chart for the pronoun distribution is shown in Figure 4.8. The

[9] We treat **bn** as a written genre rather than a spoken one

	tw	**swbd**	**OntoNotes**				
Feature			**tc**	**bc**	**bn**	**nw**	**wb**
# of tokens	43K	248K	103K	147K	225K	355K	71K
# of Personal PRPs	3082	4817	5240	3161	1594	786	1670
# of 1st Person PRPs	1071	1923	2142	1219	458	134	628
# of 2nd Person PRPs	709	1056	1083	595	175	37	195
# of 3rd Person PRPs	1298	1836	2013	1345	958	612	834
# of Possessive PRPs	612	380	343	449	403	344	519
# of 1st Person Poss. PRPs	174	185	146	151	68	28	115
# of 2nd Person Poss. PRPs	162	62	74	69	24	9	41
# of 3rd Person Poss. PRPs	255	105	122	229	311	307	359
# of Noun Phrases LS	6106	4327	4144	4959	5663	5073	4764
# of Noun Phrases SS	7610	5542	4883	7288	8667	8622	7948
# of Demonstrative PRPs	310	613	431	331	124	45	135
# of Personal PRPs in PRNs	37	671	533	162	40	17	19
# of 1st Person PRPs in PRNs	18	184	159	64	17	2	3
# of 2nd Person PRPs in PRNs	9	473	362	86	16	1	7
# of 3rd Person PRPs in PRNs	10	14	15	13	8	15	15
# of Possessive PRPs in PRNs	3	1	3	1	1	1	5
# of Noun Phrases in PRNs (LS)	57	16	16	30	17	26	28
# of Noun Phrases in PRNs (SS)	85	18	20	38	22	36	40
Nominal Density (SS) (%)	26.6	25.9	25.2	26.1	25.6	23.5	23.9

Table 4.1: Frequencies of pronouns and NPs per 40K tokens (normalized values) (PRP: Pronoun, NP: Noun Phrase, LS: Large Span, SS: Short Span, PRN: Parenthetical Clause. Possessive pronouns are not included in the count of personal pronouns. Instead, we show their frequencies in separate rows in the table. Personal pronouns are not included in the noun phrase frequencies.)

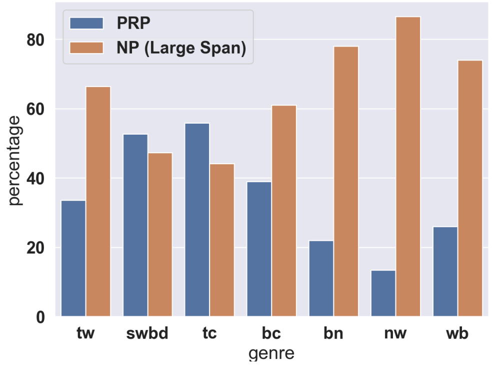


Figure 4.2: Distribution of Large Span NPs and PRPs (Nominal expressions in parentheticals are included)

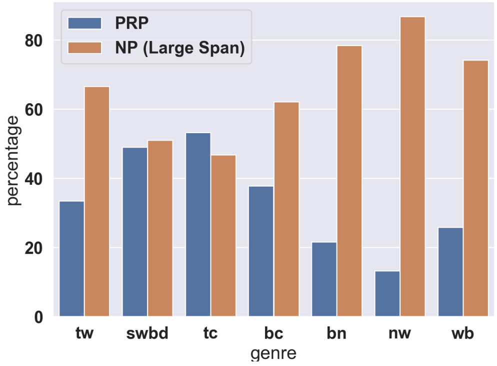


Figure 4.3: Distribution of Large Span NPs and PRPs (Nominal expressions in parentheticals are excluded)

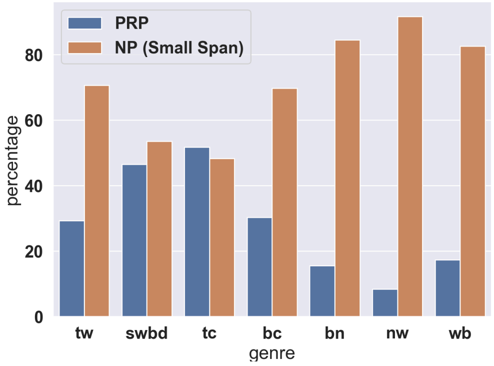


Figure 4.4: Distribution of Short Span NPs and PRPs (Nominal expressions in parentheticals are included)

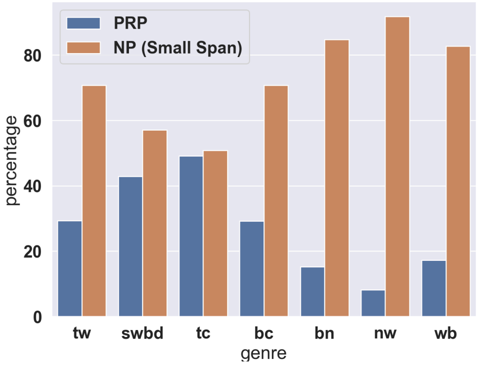


Figure 4.5: Distribution of Short Span NPs and PRPs (Nominal expressions in parentheticals are excluded)

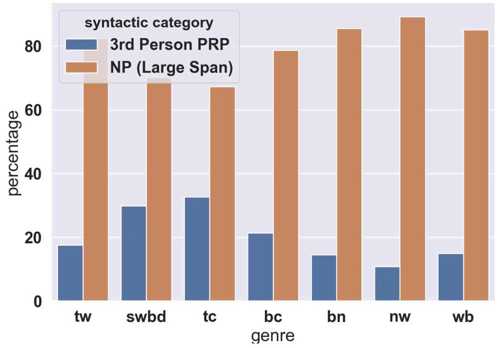


Figure 4.6: 3rd Person Pronouns vs non-pronominal NPs (LS)

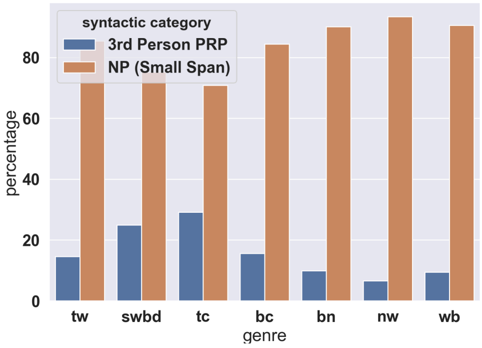


Figure 4.7: 3rd Person Pronouns vs non-pronominal NPs (SS)

same distribution without the pronouns in parenthetical clauses is shown in Figure 4.9. The significance tests indicate that the distributional PRP differences across genres are significant as shown in Table A.2. The only exception is “swbd-tc” pair for which the difference was not proved to be statistically significant. Both figures indicate similar patterns for spoken-written contrast. Conversational genres are characterized by the more frequent usage of second person pronouns than written genres.

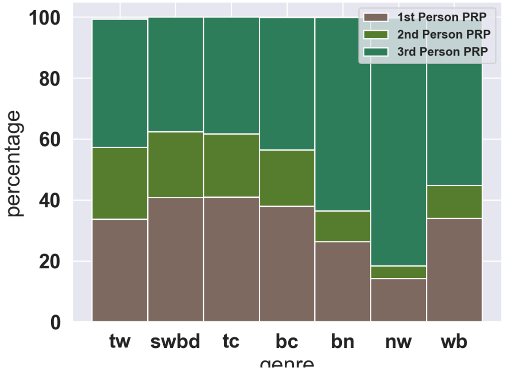


Figure 4.8: Distribution of Personal Pronouns (Pronouns in parentheticals are included)

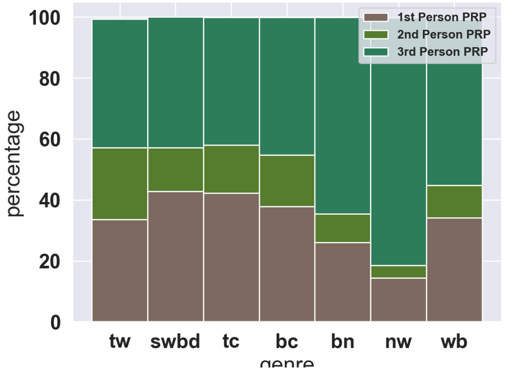


Figure 4.9: Distribution of Personal Pronouns (Pronouns in parentheticals are excluded)

4.5.2 Heaviness-based Features

Description A variety of definitions for the heaviness of noun phrases has been proposed in the literature. Wasow [1997] classifies these definitions into two groups. The first group contains the *categorical* definitions relying on, for instance, the type of nodes dominated, or the givenness of the constituents involved. The second group is composed of *graded* measures such as number of words included, or nodes/phrasal nodes dominated. Wasow compares these measures in the context of constituency ordering and concludes that graded measures are more descriptive in that context and they all work well according to the corpus-based evidence presented in the study. As Wasow’s analysis indicates that number of words in NPs is sufficiently robust to evaluate the heaviness, we use a slightly modified version of this metric and consider the number of tokens in noun phrases (i.e., NP-Length) as the measure of heaviness of NPs. In addition to the length, we also considered the number of nodes in NP parse trees (i.e., NP-Height) as a second measure. We compute the heaviness metrics for both large and short span NPs, and we exclude personal pronouns in heaviness-based comparison of NPs.

The length of an NP is the number of tokens in an NP span (i.e., number of leaves in the NP phrase structure tree). As shown in Figure 4.10, the length of the phrase, “an invasion of the privacy”, is 5 because it contains 5 tokens in the NP span (Figure 4.10.a). As for the shortest NP span examples (Figure 4.10.b), the lengths of the phrases, “an invasion” and “the privacy”, are both equal to 2.

The height of an NP is the maximum number of levels (i.e., number of nodes) in the NP phrase structure tree. As shown in Figure 4.10, the height of the phrase, “an invasion of the privacy”, is 5 because it contains 5 levels in the branch from the root node “NP” to the leaf “privacy” (Figure 4.10.a). As for the shortest NP span examples (Figure 4.10.b), the height of the phrases, “an invasion” and “the privacy”, are both equal to 3.

Thanks to the existing gold annotations of phrase constituents in OntoNotes and Switchboard, we compute NP-length and height values with an automated pipeline for those datasets. However, as mentioned in Section 4.4.3, the text spans of NPs are needed to be manually corrected for Twitter texts. In order to compute the NP-height values for

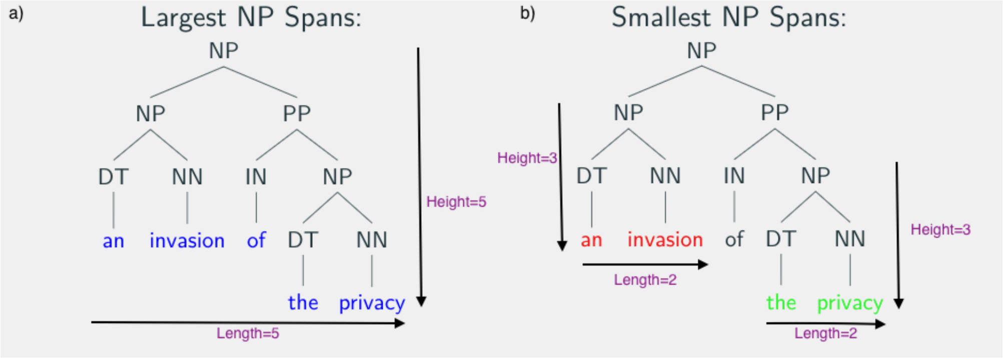
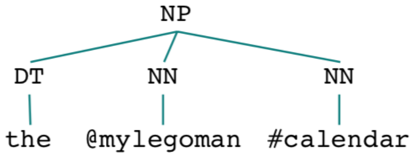


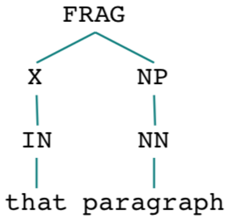
Figure 4.10: Length and Height of NP text spans

Twitter, we create automatic NP parse trees on these manually corrected NP boundaries. We do not apply manual verification to the parse trees for this measure. However, to improve the reliability of NP-height metric, we only consider the noun phrases which are captured as an NP by the automatic parser as illustrated in 4.21 (i.e., some of the NPs are parsed as fragments or other type of structures as in 4.22).

(4.21) NP-Length = 3 / NP-Height=3



(4.22) NP-Length = 2 / NP-Height=Not Applicable



Results

NP-Length We compute the average NP-length for both the largest and shortest NP spans. The results are shown in Table 4.2 and visualized in Figure 4.11.

The NP-length data do not follow the normal distribution. Therefore, we apply a non-parametric statistical test (Kruskal-Wallis test) to assess the significance of differences among genres. For pairwise comparison of the genres, we apply Wilcoxon rank sum test. The significance levels for each genre pair are shown in Table A.3.

The average NP-length values given in Table 4.2 indicate that the length of the nominal expressions for both the largest and shortest spans follows a similar pattern. This pattern

Feature	tw	swbd	OntoNotes				
			tc	bc	bn	nw	wb
NP-Length (Large Span)	2.75	3.41	2.91	4.15	4.40	5.44	5.27
NP-Length (Short Span)	1.78	1.93	1.89	2.06	2.16	2.37	2.21
NP-Height (Large Span)	3.61	4.07	3.74	4.33	4.39	4.64	4.67
NP-Height (Short Span)	3.11	3.06	3.06	3.07	3.09	3.13	3.09

Table 4.2: Average NP-Length and NP-Height across datasets

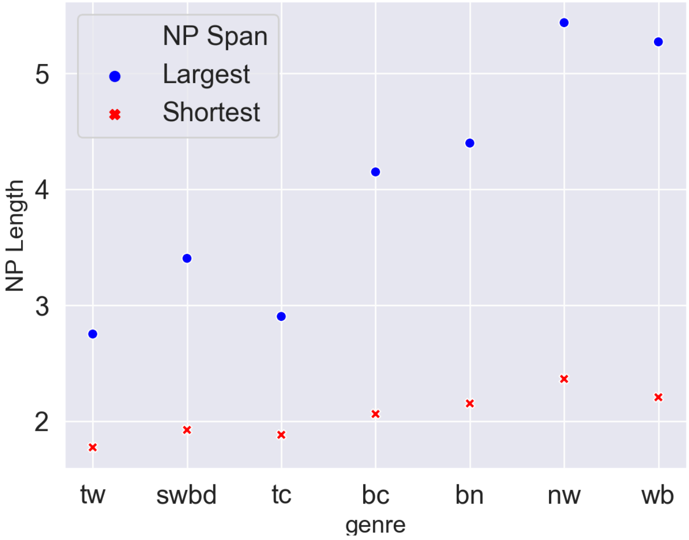


Figure 4.11: Average NP-Length

indicates that the average length of NPs is longer in written genres than in spoken genres. Twitter has the shortest NP-length which can partly be associated the character constraint on Twitter messages. The statistical significance tests indicate that these differences are not due to chance. The pairwise tests reveal the fine grained distinctions between genres. As shown in Table A.3, the tests indicate that differences in average NP-Length values are statistically significant for all the genre pairs except *tw-tc* for large span NPs and *bn-wb* pair for short span NPs (p-value<0.05).

NP-Height We compute the average NP-height for both the largest and shortest NP spans. The results are shown in Table 4.2 and visualized in Figure 4.12.

Similar to NP-length, distribution of the NP-height data does not follow the normal distribution. Therefore, we again apply non-parametric Kruskal-Wallis test and Wilcoxon rank sum test for pairwise comparison. The significance levels for each genre pair are shown in Table A.4. The differences in average NP-Height values are statistically significant for all the pairs except *tw-tc* for large span NPs, and *swbd-tc* and *bn-wb* pairs for short span NPs (p-value<0.05)

The average NP-height values given in Table 4.2 show that the height of the nominal expressions for both the largest and shortest spans follows a similar pattern. This observed pattern denotes that the average height of the NPs is greater in written genres than in spoken genres. The pairwise statistical tests reveal that large span NPs can be a more

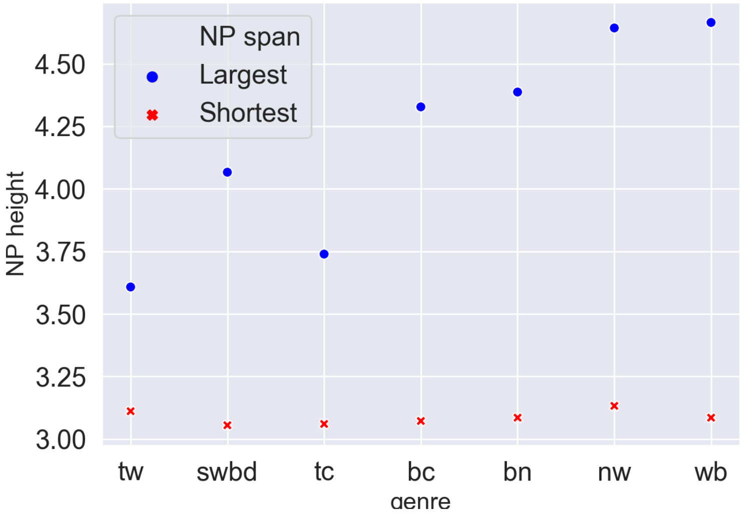


Figure 4.12: Average NP-Height

distinctive feature than short span NPs for differentiating the genres.

4.5.3 Distance-Based Features

We measure the linear distance between anaphoric 3rd person pronouns and their antecedents in terms of tokens, clauses and noun phrases. We exclude 1st and 2nd person pronouns, and non-anaphoric intensifier self-forms of 3rd person pronouns (examples given in 4.23 and 4.24) in distance-based computations. The average distance values for all the genres are shown in Table 4.3 and are discussed in the sections that follow.

(4.23) The prisoner **himself** can come to the point .. (swbd)

(4.24) .. the people around the President and surprisingly toward Bush **himself**. (bc)

A qualitative investigation of long distance anaphor-antecedent links in Switchboard indicates that long anaphoric distances can arise from the missed antecedents or wrong matching of pairs for that corpus. For instance, in example (4.25, there are 1699 tokens between two instances of “they”, but an additional mention for that entity -*channel thirteen*- was mistakenly not annotated). Another similar example is given in 4.26, where the observed distance between two instances of *they* is 879 tokens because another mention belonging the same chain, *Israel*, is not annotated. To get rid of the potential side effects of the misleading annotations, we did not take into consideration anaphoric distances that are longer than 500 tokens, 100 clauses or 150 NPs in Switchboard.

(4.25) [**they**] mention sulfur and carbon dioxide a lot [..] and *channel thirteen* [**they**]’re really um emphasizing the problem with acid rain.

(4.26) The Israelis can do anything [**they**] want [..] I was really surprised that *Israel* stayed out of it as much as [**they**] did.

	tw	swbd	OntoNotes				
Feature			**tc**	**bc**	**bn**	**nw**	**wb**
TBD	16.10	18.44	16.90	16.21	13.01	14.07	13.23
TBD′	15.97	17.38	15.65	15.85	12.98	14.07	13.21
CBD	2.85	3.35	2.97	2.44	1.66	1.45	1.42
CBD′	2.82	3.08	2.76	2.36	1.63	1.42	1.40
NBD	3.23	3.98	4.39	4.25	3.34	3.33	3.21
NBD′	3.19	3.68	4.18	4.17	3.32	3.33	3.21

Table 4.3: Average distance measures across datasets

4.5.3.1 Token-Based Distance

Description We count the number of tokens between the initial tokens of two mentions to calculate the linear token-based distance (TBD) between them. As mentioned in Section 4.3, the hashtags and usernames that are not automatically inserted by the Twitter's user interface, as well as emojis, links, and smileys are considered as single tokens in the **tw** corpus. Other tokens in all genres are compatible with PTB conventions. Discourse markers such as the fillers "um", "uhm", "well" are frequent in spoken genres (for statistics, see Table 3.1). To see the impact of these tokens on TBD, we additionally measured the distance without considering discourse markers (TBD′).

Several examples regarding how we compute TBD in different cases are shown in Table 4.4.

String[10]	TBD
[**The** manager]$_{\mathrm{i}}$ is being treated in the hospital , but [**his**]$_{\mathrm{i}}$ life ..	10
I had [**one** client who said that [**he**]$_{\mathrm{i}}$'d pay me]$_{\mathrm{i}}$ ..	5
When I saw [**him**]$_{\mathrm{i}}$ talking about [**himself**]$_{\mathrm{i}}$ ..	3
He threatened [**her**]$_{\mathrm{i}}$ [**she**]$_{\mathrm{i}}$ said .	1
He threatened her [**she**]$_{\mathrm{i}}$ said . He threatened to kill [**her**]$_{\mathrm{i}}$ ..	7
user1[11]: .. [**he**]$_{\mathrm{i}}$ is spreading . user2 : @user1 Well [**he**]$_{\mathrm{i}}$ ai n't wrong.	5[12]
Er, [they]$_{\mathrm{i}}$ [themselves]$_{\mathrm{i}}$ claim that ..	ignored

Table 4.4: Examples on the computation of token-based distances

Results The average TBD between 3rd person pronouns and their antecedents for all genres are shown numerically in Table 4.3 and visualized in Figure 4.13.

The distance data do not follow the normal distribution. Therefore, similar to heaviness-based comparison, we applied the non-parametric Kruskal-Wallis test to assess the significance of differences among genres and the Wilcoxon rank sum test for pairwise distance-based comparison of the genres. The statistical tests indicate that the differences among genres in terms of TBD values (for both settings) can be due to chance.

The significance levels of variation for each genre pair are shown in Table A.5. As expected, exclusion of discourse markers has more impact on spoken genres than written

[10] We use subscripts to mark the coreferent mentions, and **boldface** for the beginnings of mentions.

[11] The usernames are anonymized.

[12] Auto-inserted usernames are not counted as tokens.

genres. After the DMs are excluded, the pairwise differences become potentially less significant according to Table A.5. The applied non-parametric Kruskal-Wallis test to mode differences does not indicate statistical significance between production modes (p-value>0.05).

Although we do not observe statistically significant patterns according to the mode, the values in Table 4.3 designate that spoken genres (swbd, tc, bc) have longer average TBD values than written genres. Twitter conversations are closer to spoken genres in relation to the average TBD values. However, since these differences are not proved to be statistically significant, token-based distance is not a reliable measure to distinguish the production mode and the text genre.

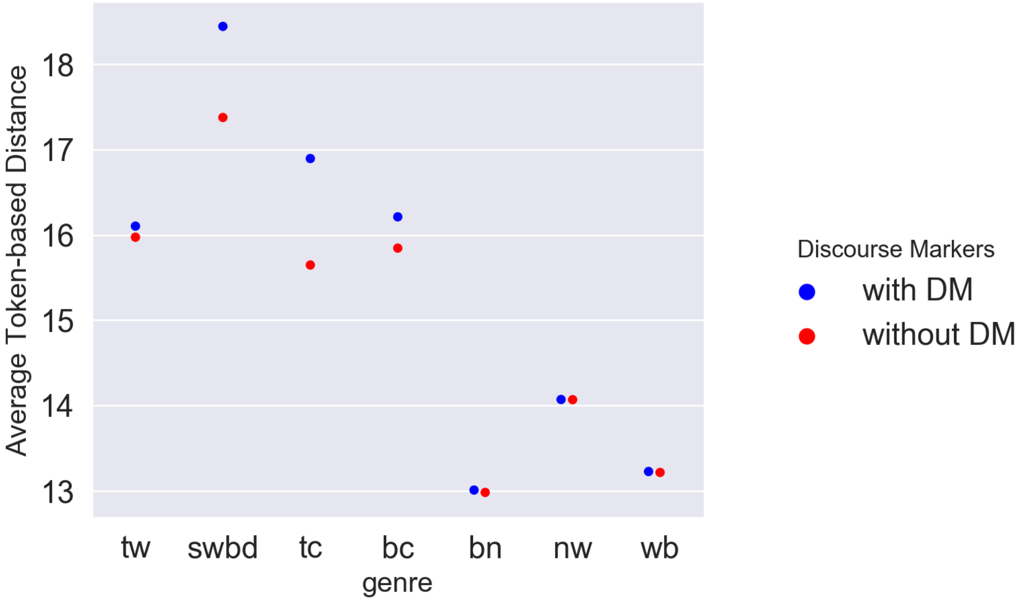


Figure 4.13: Average token-based distance across datasets

4.5.3.2 Clause-Based Distance

Description The first step in measuring the clause-based distance (CBD) between two mentions is determining the clause boundaries in the texts. We use the constituency parse trees to detect the clause boundaries. We manually correct the identified clause boundaries for **tw** corpus due to low accuracy of automatically created parse trees. We mark the first token of each clause as an indicator of a new clause and count the number of clause indicators between two mentions for calculating the CBD. The labels indicating complete clauses in the data are S, SBAR, SQ, SINV, and SBARQ; incomplete clause labels are RRC and FRAG. We took these labels as potential indicators of the beginning of new clauses. Some examples regarding how we deal with different cases in the data are presented in Table 4.5.

There exist some utterance parses in the corpus that do not have a clausal label. These utterances are considered as clauses in our calculations if they contain a nominal tag, such as “The first time?”, “Question for you.”, “The look in her eyes”, and “Pat Fitzgerald”, and/or a verbal tag, such as “Go in not go in or go in with greater strength”, “Look”, “Paid or unpaid?” and “be over”. We exclude utterances with no clause marker and with no nominal or verbal tag, such as “Hello”, “Absolutely!”, “Hopeless and angry”, “No?”, and “LOL”.

pseudo-parse[15]	# of clauses
SBARQ(**What** kind of memory?)	1
S(S(**This** is him) , S(**thank** you all for watching).)	2
S(S(**He** threatened her) she said.)	2
SBARQ(**Well** , what exactly SQ(was this incident)?)	1
S(**and** she said SBAR(**that** um S(she feels S(**she** was brainwashed))).)	3
S(SBAR(**What** S(**you** are interested in) is SBAR(**exactly** what S(we will be focusing on))).)	3
FRAG(**For** instance perhaps the chapter seven resolution.)	1

Table 4.5: Examples on the computation of clause boundaries

In **tw**, automatically inserted @-usernames are not considered as part of the clauses. Common expressions in **tw** texts such as "LOL", "haha" and emojis, hashtags are themselves not clauses, but they can be a part if they are used at the end of a clause as in "he said that lol" or they are part of the syntactic structure as in "this doesn't pass the #smelltest". Emojis, smileys, hashtags, links, final punctuation, and usernames which are not automatically inserted to the replies are included in clause spans. In examples 4.27-4.30, we demonstrate a number of clause instances in brackets from the TwiConv corpus.

(4.27) @username1[13] @username2 [Then what's it got to do with the translator ?]

(4.28) @username1 @username2 @username3 @username4 [13 is the best blur album en of :)]

(4.29) [@JoeNBC just said twice [the @washingtonpost deleted his unnamed quotes from someone around Trump]]

(4.30) [You deflect any actual critique] [because 'derp you just a fake acc'][14]

Distribution of utterances that we do not consider as clauses is presented in Table 3.1. In OntoNotes, the biggest portion of these utterances is encountered in the **tc** data (83%). Similar to **tc** where proportion of ignored vs counted utterance is 18%, Switchboard also has a high percentage of non-clausal utterances with 21% proportion of ignored vs counted utterance. In Twitter, only 6% of the utterances are not considered as clauses. As shown in Table 4.1, parenthetical clauses (PRNs) are frequent in spoken genres. To see their impact on CBD, we additionally measured the distance without considering parentheticals (CBD′).

Results The average CBD values between 3rd person pronouns and their antecedents across datasets are shown in Table 4.3 and visualized in Figure 4.14. The statistical significance of the differences for each genre pair is shown in Table A.6.

The values in Table 4.3 designate that spoken genres (swbd, tc, bc) have longer average CBD values than written genres and **tw** texts are closer to spoken genres in relation to average CBD values, for both with and without parenthetical clauses. The statistical

[13] Usernames are anonymized.

[14] Missing copular verbs are common in Twitter clauses.

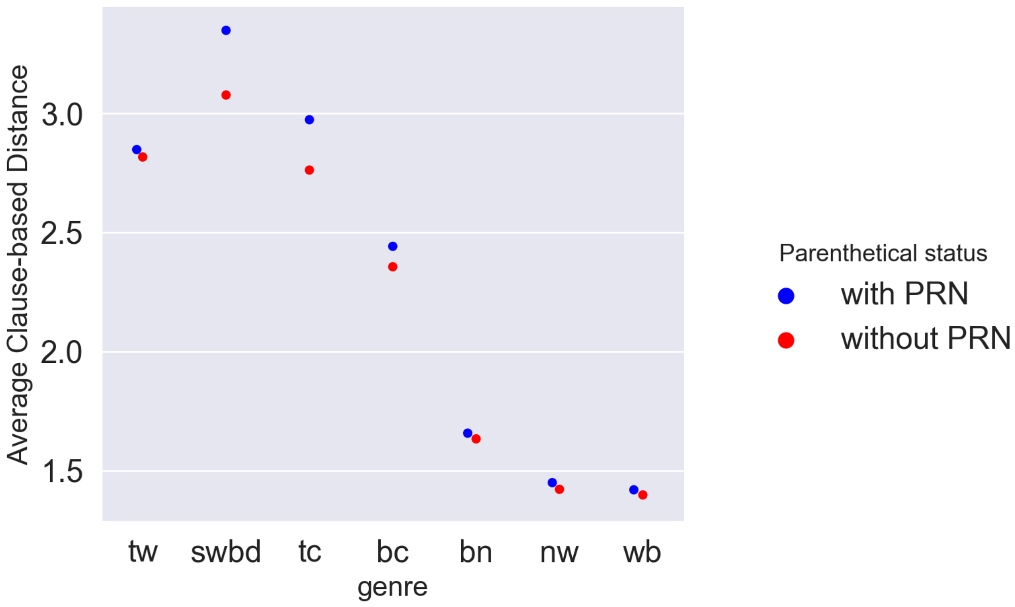


Figure 4.14: Average clause-based distance across datasets

significance tests for CBD indicate that differences between **tw** and spoken genres (*tw-tc* for CBD, *tw-tc* and *tw-swbd* for CBD′) and between two written genres (*nw-wb*) can be due to chance. Apart from those, all genre differences in terms of CBD are statistically significant (p-value<0.05) for both settings. The differences between spoken and written texts[16] are statistically significant, but the difference between spoken and Twitter texts can be due to chance (p-value>0.05).

4.5.3.3 NP-Based Distance

Description We count the number of (short span) nominal expressions between two mentions to calculate the linear NP-based distance (NBD) between them. All the nominal expressions including PRPs are considered in the calculation of NBD. We also compute the NBD without considering NPs in parentheticals (NBD′).

A number of examples regarding how we deal with different cases in the data are given in Table 4.6.

Results The average NBD values between 3rd person pronouns and their antecedents across datasets are shown in Table 4.3 and visualized in Figure 4.15.

The statistical significance levels for each genre pair are shown in Table A.7. For NBD, except from the *bn-wb*, *nw-wb* pairs, all genre differences are statistically significant. However, when nominals in parentheticals are excluded (i.e., for NBD′), the differences in *bc-nw* and *tw-swbd* pairs become not significant.

[16] Only *nw* is considered as the written genre in this computation.

String	NBD
[**The** manager]$_i$ is being treated in **the hospital** , but [**his**]$_i$ life ..	1
.. with [his mother]$_i$, **I** 'm sure **this** is why mhm at **Lisa's wedding** and, yeah, **we** were just talking about **you** and [she]$_i$ said ..	5
.. thats not true but ok lol [GAGA]$_i$ is unbothered [she]$_i$ doesnt need 15 writers..	0
user1: Oh dear, hope [your Auntie]$_i$ picks up soon xx **#findHer** user2: @user1 **It** is cos [she]$_i$ does n't want to go home lol :)	2

Table 4.6: Examples on the computation of NP-based distances (indices mark the coreferent mentions, and **boldface** the beginnings of mentions)

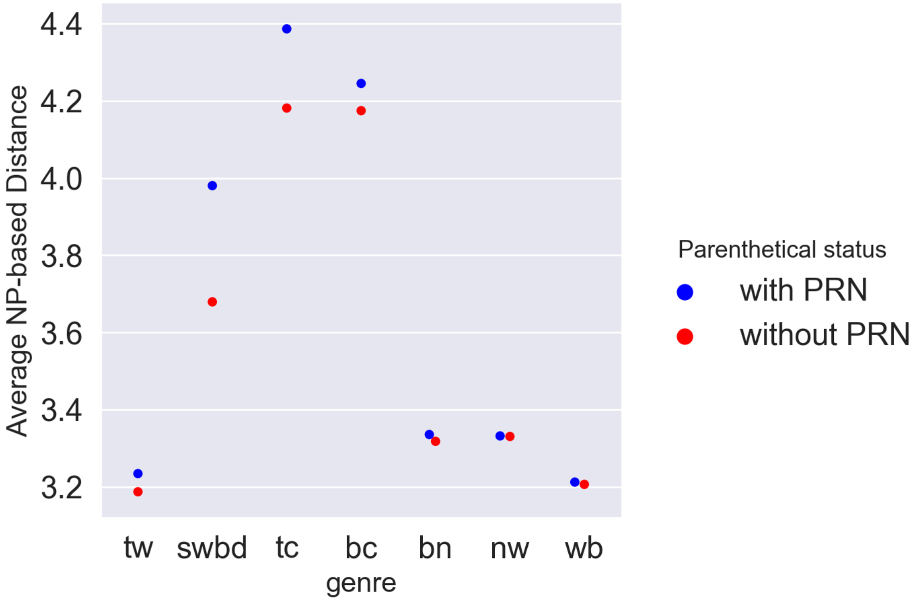


Figure 4.15: Average NP-based distance across datasets

According to the values in Table 4.3, spoken genres have longer average NP-based distances than written genres and this difference is confirmed to be not due to chance by statistical significance tests; **tw** is closer to written genres in terms of average NBD value. We calculate the density of nominal expressions for all the corpora as shown in Table 4.1 by the "Nominal Density (SS)" row. This value represents the percentage of all nominal expressions (i.e., sum of all short span non-pronominal NPs and personal pronouns) with respect to all tokens in each corpus. The nominals in parenthetical clauses are included in this calculation. The graphical representation of nominal density is shown in Figure 4.16. The figures suggest no correlation between average NBD and nominal density values. For instance, the nominal density value for **tw** is significantly higher than all the genres except from **bc** (see column 3 in Table A.7), but we do not observe a similar pattern for average NBD values.

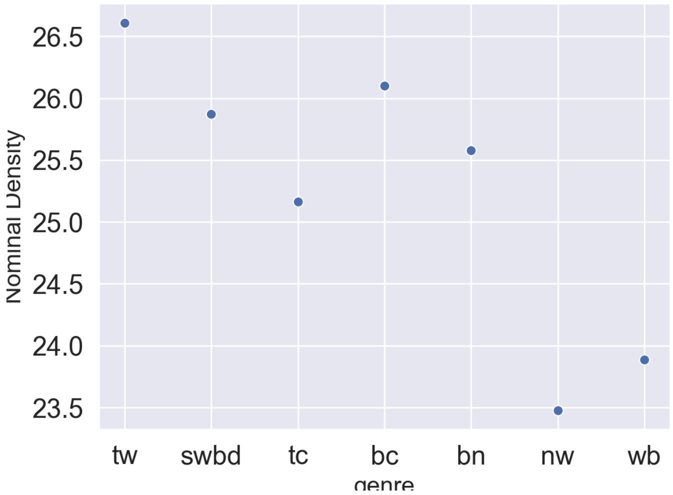


Figure 4.16: Nominal density across datasets

4.6 Discussion

Following the notion of spoken-written continuum proposed by a number of linguists (e.g., Tannen [1982a],Koch and Oesterreicher [1985],Biber [1988]), we do not assume a binary distinction between the language modes but a spectrum of genres, which only loosely corresponds to the two modes. In OntoNotes, for example, the genre of “broadcast news” contains edited speech that differs in many ways from the spontaneous speech of “telephone conversations”.

Our analyses in the previous sections lead to rankings of the genres, which collectively suggest a general pattern. We observe that two conversational spoken datasets, swbd and tc, and two written genres, nw and wb, are always located closely (if not adjacent) in the ranking of genres in terms of the average values of the features investigated, and these two pairs are situated at the opposite ends. The first pair (i.e., swbd/tc) is composed of the datasets of telephone conversations while the second pair (i.e., nw/wb) contains the texts of news and web blogs. Members of those pairs (i.e., news and telephone conversations) correspond to what [Fox, 1987, p.138] refers to as unmarked text-types of respective modes, which indicate what one typically does when writing and speaking.

Our frequency-based analysis shows that relative frequencies of pronouns and NPs in the swbd/tc pair are close to each other, whereas in nw/wb, NPs are substantially more frequent. Our results indicate that pronoun usage is more common in spoken conversational genres (swbd, tc, and bc) than in written genres, which is consistent with Fox’s findings. Fox also argues that pronouns are more frequent than noun phrases in both modes, which is in conflict with our finding that pronouns are less frequent than noun phrases in almost all genres.

Jonsson [2016] analyzes the language in computer-mediated communication (CMC) using the multi-dimensional analysis proposed by Biber [1988]. Jonsson examines both synchronous and asynchronous CMC (*ACMC*) and contrasts their linguistic characteristics with *Speech* (e.g., telephone conversations, prepared speeches etc.) and *Writing* (e.g., personal letters, academic texts etc.). Twitter is included under her classification of asynchronous CMC along with SMS, e-mails, and others. One of the features used in her analysis is the frequency distribution of personal pronouns. We integrate her results into

Table 4.7, which also shows pronoun frequencies for the datasets we examine. The frequencies in Jonsson's study are normalized per 1,000 words, whereas in our datasets per 1,000 tokens. Therefore, the numbers in the first three columns are not directly comparable to the rest of the table. We can nevertheless discuss the patterns observed in two studies. Jonsson [2016] finds that all forms of CMC include more first person pronouns than *Speech* or *Writing*. In addition, usage of third person pronouns in *ACMC* is less frequent than in both *Speech* and *Writing*. Our findings on first- and third-person pronoun frequencies differ from those of Jonsson's findings because we observe that conversational Twitter data is placed between the spoken and written genres for both pronoun types. On the other hand, our findings support Jonsson's observations on the use of second person pronouns, showing that both **tw** and **ACMC** frequencies in Table 4.7 are situated between spoken and written texts in terms of the frequency of second person pronouns, though the values are closer to the spoken side.

	Writing	*Speech*	*ACMC*	**tw**	**swbd**	**OntoNotes**				
Feature						**tc**	**bc**	**bn**	**nw**	**wb**
1st pers. pro.	17	52.8	57.8	31.1	52.7	57.2	34.2	13.2	4.1	18.6
2nd pers. pro.	5	23	17.6	21.8	27.9	28.9	16.6	5	1.2	5.9
3rd pers. pro.	30.7	29.2	26.9	38.81	48.5	53.4	39.3	31.7	23	29.9

Table 4.7: Frequencies of first, second and third person pronouns (normalized values): The first three columns, *Writing*, *Speech* and *ACMC* from *Table 4.2* in [Jonsson, 2016, p.122] are normalized per 1,000 words, while the other columns are normalized per 1,000 tokens. Both Jonsson's and our statistics include possessive pronouns.

A comparison of distance-based features indicates that the textual distances between anaphoric pronouns and their antecedents are longer in swbd/tc than nw/wb. This result is in line with Fox [1987]'s finding concerning the clause-based distance but contradicts with Biber [1992]'s findings related to NP-based distance, which suggests the opposite pattern. The reason for this difference could be the dissimilarities between the genres investigated. For instance, we do not observe the same pattern for the genre of broadcast news, although it is produced in spoken mode. And lastly, nw/wb genres contain heavier NPs than swbd/tc do, which is in line with Amoia et al. [2012]'s findings. In our analysis, NP-length in **tw** even drops below the level of spoken genres.

The differences between the opposite ends of the mode spectrum are statistically significant for all the measures except token-based distance. The significance levels can differ when parenthetical clauses and discourse markers are excluded from the computation of features, but these differences do not affect the statistical significance of differences between the swbd/tc and nw/wb pairs. The placement of the **bc** and **tw** genres in our genre ranking differs according to the measure. For instance, **tw** is located at the spoken end of the spectrum with respect to NP-Length, but close to the written end with respect to NP-based distance. For the other features, it is situated between the spoken and written ends, typically being closer to the spoken end in terms of the quantitative values of the metrics.

In addition to rankings derived from average-value orderings, we also use **hierarchical clustering** for grouping the genres based on the quantitative features. To obtain the clusters, we first normalize the data size, and use the complete linkage method for hierarchical clustering (i.e., the maximum distance between the members of clusters is taken as the cluster distance). For this purpose, we apply the *hclust* method in R with default settings. We provide the genre clusters in Appendix B. Similar to the observations mentioned above, the swbd/tc and nw/wb pairs are grouped together for all the metrics except NP-based distance for swbd/tc (see in Figure B.5).[17] Again, for **bc** and **tw**, the groupings vary from feature to feature. For instance, in Figure 4.17, **bc** is clustered in the same branch with nw/wb, whereas in Figure 4.18, it is grouped together with the swbd/tc pair.

The ranking-based and cluster-based groupings do not always overlap (as in the NP-based distance case), but the proximity of the swbd/tc and nw/wb pairs is observed as a common pattern in both cases. A few exceptional cases of this pattern require more attention. The groups of **bc** and **tw** depend on features. The genre of **bn** is consistently close to the nw/wb pair, which is not unexpected given the edited content of the broadcast news. We consider these findings (i.e., existence of more rigid clusters as well as the floating genres between the groups) as supportive evidence for considering spoken and written modes as a continuum rather than discrete concepts. It turns out that Twitter conversations occasionally exhibit both spoken- or written-like characteristics. Twitter conversations, however, occupy a medium position in the continuum for the majority of the features we investigated in this study. Its exact position differs from feature to feature.

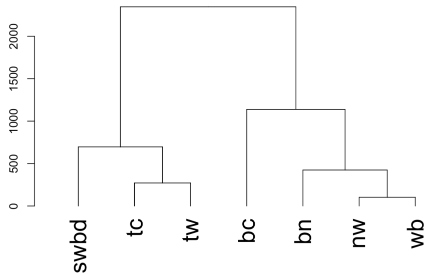


Figure 4.17: Genre clustering based on CBD

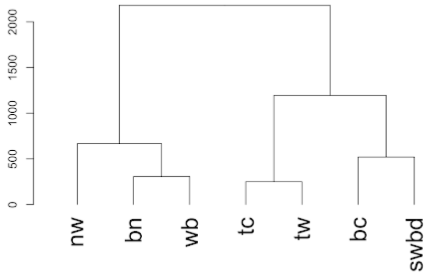


Figure 4.18: Genre clustering based on NP-Length (LS)

4.7 Conclusion

We present an in-depth study of coreference patterns across genres that involve different production modes (spoken and written). As a prerequisite, we harmonize the annotations of three corpora, and to our knowledge, this is the first systematic comparison of its kind. Our findings and interpretation for the spoken–written dichotomy and the placement of Twitter conversations are given in the previous section. Some of the features used in this study such as pronoun and NP frequency, frequency distribution of personal pronouns, and the length of noun phrases are also considered in more general discussion of spoken and written language (e.g., [Jonsson, 2016]) and our results are partly compatible with these previous findings. We also discuss in the previous section the incompatibilities with the prior research on variation in coreference patterns.

In Section 2.3.2, we discuss a number of theoretical approaches that associate linguistic features of the texts with certain dimensions, such as “involvement” vs “detachment” [Chafe, 1982] or language of “immediacy” vs “distance” [Koch and Oesterreicher, 2012]. In our work, although we find these discussions about fine-grained characteristics of genres and modes interesting and important, given the scope of this study, we leave them as a future effort.

The differences between the anaphoric distances are more pronounced between the genres on the different extremes of the spoken-written continuum. The variation in anaphoric

[17] We do not make clustering according to NP-Height because NP-Length is a more reliable metric of heaviness due to manually corrected NP spans.

distance therefore appears to be more influenced by the genre than the mode. In order to examine whether the mode has an effect at all on the anaphoric distance, we conduct a story continuation experiment that is complementary to corpus analysis. The findings of this study will be presented in Chapter 5. Finally, in addition to the descriptive analysis on genre differences, we also see our results as potentially fruitful for adapting automatic coreference resolution to genres and modes, in the light of small amounts of training data, where knowledge about the differences in patterns could be utilized.

5

Variation in Coreference in Spoken and Written Language: Experimental Evidence

We are investigating the variation in spoken and written English in terms of coreference-related features of language. One of the most studied aspects of coreference is the textual distance, either linear or hierarchic, between the referential expressions denoting the same entity (henceforth, anaphoric distance). In this chapter, we aim to expand corpus-based research on anaphoric distance with experimental evidence.

The distance between anaphors and their antecedents (the closest previous mention of the same referent) is considered as one of the most important factors influencing the referential choice [Ariel, 1990, Chafe, 1976, 1994, Garrod and Sanford, 1982, Givón, 1983]. Experimental evidence indicates that anaphoric distance affects the comprehension effort for the resolution of referential expressions. For instance, extending the anaphoric distance in an experimental setup increases the comprehension time for the referential expressions [Clark and Sengul, 1979, Streb et al., 2004].

Similar to the other relevant comparative studies (e.g., Fox [1987], Kunz et al. [2016]), our corpus-based study in Chapter 4 demonstrates that there exist differences between spoken and written corpora with respect to the anaphoric distance between third person pronouns and their antecedents. Inline with the findings of Fox [1987], we demonstrate that the average anaphoric distance between third person pronouns and their antecedents is longer in spoken texts than in written texts.

Spoken-written comparison in terms of coreference variation shows greater contrasts for the intrinsically different genres. For instance, we compare different genres in terms of clause-based anaphoric distance in Chapter 4 and observe that the difference is more prominent for telephone conversations and news texts (2.97 vs 1.45) than for broadcast news and written news (1.66 vs 1.45). Although telephone conversations and broadcast news are both produced verbally, written news differs significantly from telephone conversations regarding spontaneity and informality, but not from broadcast news. Furthermore, we observe that full noun phrases (e.g., proper names, definite and indefinite noun phrases) are more frequent than third person pronouns for both modes, with third person pronouns being more common in spoken than in written genres.

In adddition, spoken and written texts diverge in terms of textual features. For instance, as we show in Chapter 4, certain discourse markers (e.g., mhm, oh yeah), for example, are widespread in spoken conversations but uncommon in news texts and spoken genres are more intense in terms of the number of clauses than written genres. In another comparative study, Schäpers [2009, p. 137] examines spoken and written corpora that are composed of same number of words. Schäpers also reports that spoken texts have a higher number of clauses than do written texts. Consequently, clauses in written texts are longer. In addition, Schäpers notices discrepancies between modes regarding how the clauses are linked to each other. For instance, coordination is more common in spoken language, whereas in written texts, subordinate clauses are more common.

Akinnaso [1982] argues that comparison of the modes (spoken vs written) should be conducted on the texts that are produced by the same individuals, exhibit same degree of planning, formality, interactivity (e.g., monologue or dialog) and belong to the same genre. In line with Akinnaso's argument, in this study, we have the motivation that comparing the spoken and written modes for the same genre, in a setting where the textual differences (e.g., clause length and clause types) are eliminated, can provide more insight about the impact of the mode, when both modes convey a similar level of spontaneity, informality and interactivity[1]. Therefore, in order to verify the results of the earlier corpus studies (e.g., spoken mode allows more pronoun usage in longer anaphoric distances in comparison to written mode) and get a more precise account of the distance question within a single genre, we design a story continuation experiment using a crowdsourcing platform[2]. In this study, participants are expected to produce texts, in either spoken or written form, and use referring expressions denoting the main character in the stimuli. We examine production biases in response to spoken and written stimuli, where clause-based anaphoric distance is manipulated systematically.

Our first hypothesis is that spoken responses will include more frequent use of pronouns than written responses if all noun phrases in responses are considered. Our second and main hypothesis is a specific version of the first hypothesis: spoken responses will include more frequent pronoun usage as a first reference to the main character than written responses for long stimuli. (i.e., spoken responses allow for longer anaphoric distance). In addition, we aim to provide quantitative results on various aspects of the spoken-written mode contrast.

This chapter is structured as follows: In Section 5.1, we present relevant previous work in the literature. Section 5.2 describes the design settings and execution procedures for running and evaluating the experiment. We report the results of the experiment in Section 5.3, employing descriptive and inferential statistics. In addition, we present the outcome of initial classification experiments we ran on the data in Section 5.3. In Section 5.4, we discuss our findings. Section 5.5 concludes and explores future directions.

5.1 Related Work

Various contrastive corpus-based studies investigate the differences in terms of average anaphoric distance across spoken and written modes. Since the experimental factor in our study is the linear textual distance in terms of clauses (i.e., clause-based distance), we include only the studies investigating the contrasts on clause-based anaphoric distance.

Fox [1987] examines the use of anaphoric third person singular human references (pronouns and noun phrases) in spoken and written texts. The data is composed of conversations (face-to-face and telephone conversations) and expository prose (newspaper articles and psychoanalytic biography). Fox reports that conversations contain longer clausal distances between pronouns and their antecedents than written texts do (2.52 clauses in conversations vs 1.21 in written texts). The findings, in our comparative corpus analysis presented in Chapter 4, are compatible with Fox [1987]. In our analysis, the difference in clause-based anaphoric distance is more prominent for conversational texts and news texts (2.97 vs 1.45) than for broadcast news and written news (1.66 vs 1.45), although

[1] Both spoken and written texts are produced by the same individuals. For details about the experimental procedures, see Section 5.2

[2] https://www.crowdee.com/

conversational texts and broadcast news are both produced in the spoken mode.

In addition to the anaphoric distance, in our corpus study, we compare the spoken and written genres also in terms of the relative distribution of nominal expressions according to their syntactic categories (i.e., pronouns vs non-pronominal NPs). Our findings indicate that third person pronouns are more frequent in spoken conversations than in written news (33% vs 11%). Fox [1987] compares the frequencies of pronouns and full NPs in settings with no interfering referent, with an interfering referent of a different gender, and an interfering referent of the same gender. Her findings show that in the absence of an intervening referent, pronouns are used more intensely for both environments (64% in written vs 94% in spoken language) compared to full NPs. The use for pronoun decreases when there exist interfering referents. When a referent of the same gender interferes, for example, the usage of pronouns drops to 13% in written texts and 57% in spoken texts.

In previous experimental research, the impact of anaphoric distance on pronoun processing has so far been studied within a single mode, where the anaphoric distance is modified to manipulate the working memory. Clark and Sengul [1979] demonstrate that text distance between antecedents and the anaphor affects reading time. In their experiments, participants took less time comprehending a sentence when the referents of such noun phrases were mentioned one sentence back than when they were mentioned two or three sentences back. Streb et al. [2004] conduct experiments in which they manipulated the number of intervening words between the anaphor and its antecedent, hence altering the anaphoric distance. In line with Clark and Sengul [1979], they find that comprehension time increases with the expanded distance. This effect is held for both pronouns and proper names. They also report that the cases of “medium” (antecedent in the previous sentence) and “far” (antecedent 2 sentences before) are not distinguishable.

In contrast to such studies, we examine the variation between spoken and written narrative discourse with a production experiment. To the extent of our knowledge, there is no previous work with the same goal and strategy.

5.2 Experiment

The starting point of this experiment is the assumption that heavier expressions (e.g., proper names) are produced more often than lighter ones (e.g., pronouns) when a referent is less active or salient; this has been validated by multiple studies (e.g., [Chafe, 1976, Ariel, 1990, Arnold, 1998]). The anaphoric distance is highly correlated with accessibility or salience in these studies. We here investigate the question of whether there is a difference regarding the referential choice across spoken and written modes, related to the varying length of anaphoric distance. In this experimental work, we build on our corpus-based analysis in Chapter 4 and aim to provide empirical evidence on mode distinctions in language production regarding the anaphoric distance. We hypothesize that spoken responses in our story continuation experiment will include more prominent pronoun usage than will written responses for long anaphoric distance.

The interface of this experiment is designed and implemented as a crowdsourcing study. Crowdsourcing is intensively used to generate language resources for NLP tasks, such as argument mining [Lavee et al., 2019] and coreference annotation [Poesio et al., 2019][3]. These studies and many others (e.g., [Enochson and Culbertson, 2015, Iskender et al., 2021]) report promising quality assurance results for the crowdsourced data and, hence validate the effective use of crowdsourcing even for complicated NLP tasks, such as

[3] There exist different forms of crowdsourcing, such as microworking and game-with-a-purpose [von Ahn and Dabbish, 2008]. In contrast to our study and the previous crowdsourcing studies we listed above, Poesio et al. [2019] uses the game-with-a-purpose approach, which due to the general similarity, is a form of applying crowdsourcing to coreference research.

argument annotation[4].

We prefer to use crowdsourcing instead of a laboratory setting because crowdsourcing platforms provide an efficient way for reaching the native speakers of English located in English speaking countries, from both financial and organizational perspectives. We recruit native speakers of English for our experiment to ensure a comparable level of English competency for all participants. Crowd workers participating in this study are paid according to the minimum hourly wage of the country of their residence.

Our task (i.e., story production) does not require expertise in any specific linguistic field. That being the case, in order to ensure the quality of the responses, we only needed to confirm the linguistic competence and the engagement of the participants in the study. Therefore, we recruit the participants via a qualification study (see Section 5.2.1). We create a pool of crowd workers after the evaluation of the responses collected in the qualification phase and then use that pool to recruit the participants in the actual experiment.

The experimental factor in the main experiment is the length of the short stories which are used as the stimuli for triggering the continuations. The test conditions of the experiment are constructed by manipulating the story lengths as 1, 2, 3, 5, 8 clauses. The design principles applied in the creation of the stimuli are presented in Section 5.2.3.2. In order to validate the experimental settings (e.g., stimuli design, instruction wording etc.), we conduct a pilot study before implementing the actual experiment (see Section 5.2.2). The pilot study provided practical hints for refining the experiment stimuli so that they can more effectively serve our purposes. We conduct the main experiment 8 weeks after the pilot study. The details of the main experiment are given in Section 5.2.3.

5.2.1 Participant qualification study

The qualification study is implemented as one crowdsourcing task (i.e., a set of stimuli/questions which should be completed and submitted by a crowd worker in one attempt). It is a smaller version of the actual experiment with fewer stimuli[5], yet containing additional questions for the qualification purposes. Participants of this study were expected to continue the short stories in the stimuli, which were provided in either audio or written form. In addition to the story continuation sections, comprehension questions were also inserted into the workflow in order to check the engagement of the crowd workers. Only the workers who declared being a native speaker of English were allowed to complete the qualification study.

The first page in the qualification study is the information page and contains the description in 5.1.

(5.1) **Native English speakers**: Qualify for “Story Continuation” tasks and earn up to 10 Euro today by continuing very short stories (up to 3 sentences). This is your qualification test. Duration is approx. 6-8 minutes.

Requirements: English native speakers only; Microphone and speaker or headset needed.

On the next page, we describe the experiment and get a confirmation from the crowd workers about whether they are native speakers of English. We use the instructions in 5.2 and 5.3 for this purpose.

[4] Despite its popularity, crowdsourcing platforms also get criticisms, which we acknowledge as important, concerning the data quality they provide [Fort et al., 2011] and work conditions of the crowdworkers [Webster, 2016, Panteli et al., 2020]

[5] Stimuli used in the qualification task are less restricted than the stimuli in the actual experiment as this task serve for participant qualification rather than hypothesis testing.

(5.2) Dear participants, thanks for taking part in this qualification for our **story continuation** micro task series. Here, you will see 6 very short stories in between 1 to 8 sentences of length. You are to conceive a continuation of these stories, with at least 1 and at most 3 sentences. We will present the stories in different ways to you. You are to respond to written stories by writing, and to the audible stories by short speech recordings. For every story, please try to **imagine what happens next to the main character** and write or say a continuation that comes to mind quickly, without very elaborate planning!

(5.3) We are looking for up to 100[6] English native speakers who show best results in this qualification in order to participate in up to 60 more stories. In order to participate, please confirm that you are an English native speaker or cancel this task.

If a crowd worker confirms being a native speaker of English, we then present the technical requirements for the experiment. For instance, they should allow the browser to access their microphone in order to complete the study. We insert an audio message into the page to confirm that the users can hear our audios and understand the content. The message asks for the result of a simple mathematical operation ("What is the result of 3 plus 4?"). The crowd workers are expected to write the answer in the text box provided on the same page. If the answer of the arithmetical operation was correct, we then continued with the qualification task.

The qualification study contains two comprehension and six story continuation questions, each shown on separate pages. The story continuation questions were displayed randomly, with no specific preference as to the order. Crowd workers should pass to the next page manually (by pressing the 'Next' button) after completing the request on the current page.

The story continuation questions are containing three written and three audio stories. 5.4 represent an example for a written story and 5.5 is an audio story[7].

(5.4) The door opened quietly, the dark silhouette of a man entered the room, and the door closed again. The man seemed to be looking for something.

(5.5) Jack is very happy because he got promoted at work. He is now organizing a big party to celebrate this.

The crowd workers are expected to continue to the stories by imagining "what happened next to the main character" and generate responses using the same mode as the story. For this purpose, we provide a text box for responding the written stories and a voice recording button for the audio stories. For the audio stories, we described what we expect from the participants with the statements in 5.6 and for the written stories in 5.7.

(5.6) Now go ahead, listen, try to imagine what happens next to the main character and record a continuation of up to 3 sentences that comes to mind quickly, without very elaborate planning.

(5.7) Please try to imagine what happens next to the main character again. This time, write a continuation of up to 3 sentences that comes to mind quickly! This is how the story begins [HERE THE STORY COMES].

[6] We qualified the participants in two different crowdsourcing tasks, which have the same instructions. That's why the number stated here is smaller than the target number of participants.

[7] The text-to-speech interface at https://ttsmp3.com/ was used to create audio versions of the stories in all the phases of this experimental study.

The comprehension questions are about the stimuli in 5.4 and 5.5. Therefore, they always follow these stories in the flow. 5.8 shows the comprehension question about 5.4, which asks for the best description of the situation among the given options, and 5.9 is a question about the situation in 5.5.

(5.8) Please select the answer that best matches according to your understanding.

(5.9) What made Jack to feel happy?

The crowd workers are expected to choose the most suitable option (3 options were provided for each question) to respond to the comprehension questions. We continue the qualification task even if a crowd worker did not choose the correct option but use all the answers when evaluating the qualification responses in choosing the eligible participants.

At the end of each qualification task, we ask the crowd workers to provide information on their age, gender (female, male, divers), level of education (less than high school, high school, Bachelor's, postgraduate) and spoken English variety (American, British, Other) for anonymous statistical analysis.

We went over all the responses of the qualification study manually and eliminated 19 crowd workers considered to be not eligible to participate in the main experiment. Some of these unqualified crowd workers seemed to not interpret the instructions correctly, and for the others, the audio quality of their recordings was not good. We created a pool of 556 crowd workers at the end of qualification phase.

5.2.2 Pilot Experiment

40 participants among the qualified crowd workers were recruited for a pilot experiment. These participants later also joined the actual experiment, which was conducted 8 weeks after the pilot study.

The structure of the pilot study is the same with the actual experiment. As in the qualification study, we show short stories to the participants, both in spoken and written modes and ask them to continue the stories by imagining "what happened next to the main character", using the same mode as the story. The participants see only one stimulus on the active page. The instructions are similar to the qualification study, with slight changes in the wording as demonstrated in 5.10 for written and in 5.11 for spoken story continuation collection.

(5.10) Here comes your next story beginning. Please always try to imagine what happens next to the main character! Write a continuation of up to 3 sentences that come to mind quickly. This is how the story begins [HERE COMES THE STORY]

(5.11) Please continue with listening to the next story and try to imagine what happens next to the main character! Please record a continuation of up to 3 sentences that come to mind quickly!

In the pilot study, we observe substantial differences in referential choice regarding the story length. For instance, for the written stories in 5.12 and 5.13 which differ significantly in length, we observe a great variation in the participants' responses in terms of pronoun usage.

(5.12) James went to IKEA yesterday.

(5.13) Betty wrote a new novel. It is a crime novel. The book is called Insomnia. The story passed in India, therefore it got attention from Indian media. After a review was published in newspapers, it became popular. Everybody knows it now.

83% of the continuations for 5.12 include a pronoun as the first reference to "James", whereas a pronoun is used only in 18% of the continuations for 5.13 as the first reference to "Betty". These observations are compatible with the previous studies in the literature (see Section 5.1). This outcome indicates that we replicate the previous results with our experimental design, which we consider as proper evidence for the validation of the design.

In addition to the validation of the experimental design, the pilot experiment is also beneficial for finalizing the stimuli structure of the main experiment. We discover practical hints about the stimuli design, which are useful for getting more responses about, and hence more referential expressions to, the main character. The following are the conventions that emerged from the investigation of the responses in the pilot experiment:

1. We do not use direct speech in the stories because participants tend to continue the statements similar to 5.14 by extending the included speech text instead of mentioning "what happened next to the main character". An example continuation for 5.14 is given in 5.15, where a participant is expanding Clara's speech instead of mentioning what happened to her next.

   (5.14) Clara whispered resentfully, "it is almost 3 o'clock".

   (5.15) "Where were you?"

2. We try to avoid using transfer verbs (e.g., "buy", "sell", "give") in 1- and 2-clause-length stimuli. Since these verbs introduce a new object and can shift the focus to that object, the responses to such stories are usually about the object instead of the main character. For instance, continuation examples from the pilot experiment for the story 5.16 are given in 5.17 and 5.18 below.

   In this example, in order to get more responses about "Patricia" rather than "the bicycle", we modified the story as in 5.19.

   (5.16) Patricia bought a new bicycle.

   (5.17) The new bicycle was red and fast and had a really loud bell on the front.

   (5.18) It was bright read with strings coming up the handle. It was a bike that every girl wanted.

   (5.19) Patricia rode all day yesterday.

3. We do not introduce a human entity in the stories apart from the main character (see Section 5.2.3.2 for details). Therefore, it is inevitable to insert non-human entities into long (5- and 8-clause-length) stories. In those cases, we prefer to not focus only on one object but mention different circumstances, to avoid responses about the objects. Therefore, instead of 5.20, for instance, we prefer to use 5.21 in the final stimuli set[8].

   (5.20) Mark moved to a new apartment. It has three rooms. The rooms are spacious, the kitchen is renovated. It is located in a quiet neighbourhood. It has a balcony. The view is impressive when the sun goes down over the sea."

   (5.21) Gabriella moved to a new neighbourhood. The previous area was chaotic because it was very central. This new neighbourhood is quiet. There are parks nearby. The new house is great. It is quite peaceful. A whole new life can start here.

[8] Change in the protagonist's name is only in order to balance the female-male names in the final set.

5.2.3 Main Experiment

5.2.3.1 Participants

Participants of the main experiment are recruited from the qualified crowd worker pool generated according to the outcome of the qualification study described in Section 5.2.1. Recruited participants are mostly based in US and UK. 245 English speakers, who declared to be native in English, participated in the main experiment. Mean age was 36.7 years (range 18–64) (based on the age information provided by 235 of the participants).

Distribution of participants with respect to

1. gender is 58% female, 40% male, and 2% divers (based on the gender information provided by 235 of the participants).
2. English variety is 48% American English, 50% British English and 2% other (based on the language information provided by 216 of the participants).
3. the level of education is 2% with less than a high school diploma, 34% with a high school degree, 45% with a bachelor's degree, and 19% with a postgraduate degree (based on the level of education information provided by 216 of the participants)

5.2.3.2 Materials

We consider the following features in interaction with textual anaphoric distance and restrict the structure of the stimuli in order to minimize the variation in terms of these features.[9]

1. Average clause length
   - We try to create constant length clauses in the stimuli (i.e., 5 words). However, there are cases where we need to bend this rule, to ensure naturalness in the stories. In those cases, we keep the average clause length (i.e., total number of words[10] in the story divided by the number of clauses) constant at 5 words.
2. Token types
   - Same stories are used for both audio and written stimuli in order to minimize the impact of diverging tokens and token types.[11]
3. Clause types
   - Only finite subordinate clauses are used in the stories.
   - For establishing explicit relations, only temporal and causal discourse connectives are used.
4. Bridging anaphora
   - We avoid establishing bridging anaphoric connections with the main character in the stimuli.

In addition to the above constraints, we also consider the following principles in the design of the stimuli:

[9] In addition to the features we took into account in the experimental design, there exist other potentially important features, such as the number of referring expressions in the stimulus and the length of the name of the main character in the stimulus. We leave the exploration of those features in future studies.

[10] Excluding the connectives linking the clauses.

[11] Stimuli with the same content are shown to the mutually exclusive groups, so each participant is exposed to either audio or the written version of the same story.

1. The experimental factor in this study is the length of the stories. The test conditions of the experiment are constructed by manipulating the story lengths as **1, 2, 3, 5, 8 clauses**.

2. The main character is referred only once in each story, always introduced in the first clause and referred to by a proper name.

3. Only one human reference (main character) is introduced in the stimuli.

4. Stories are always written in past tense.

5. Stories always start with the main character in subject position in order to avoid possible complications due to the differences in syntactic roles.

6. 8 stimuli for each clause length condition (4 with common male names + 4 with common female names) are created. As a result, we have 40 stimuli in total.

7. 20 filler stories having no constraints are used in addition to the 40 actual stimuli. Responses to the filler stories are not included in the analysis.

The above-mentioned constraints and principles regarding the stimuli design are determined at the beginning of the design phase. The stimuli set is finalized after evaluation of the pilot experiment described in Section 5.2.2. In Table 5.1, we present sample stories from the final set for each test condition and filler stories.

Condition	Example
1-clause length	Robert finally solved the mystery.
2-clause length	Barbara had a quick lunch before the doorbell rang once again.
3-clause length	Lisa washed the dishes resentfully after the party ended. It was a long and gloomy night.
5-clause length	Carol made a peppermint tea before the sun rose. Herbs can balance the emotions. Peppermint helps for anxiety so it was a perfect choice for this morning.
8-clause length	Charlotte was tired. There are good days in life; there are also challenging days. Today was one of the latter. There was nobody around. The building was almost empty. It was getting dark. Even the computer turned to sleep mode.
Filler	The door opened quietly, the dark silhouette of a man entered the room, and the door closed again. The man seemed to be looking for something.
Filler	The evening passed quietly, unmarked by anything extraordinary. At night, Darcy opened his heart to Jane.

Table 5.1: Stimuli examples

5.2.3.3 Procedures

As previously stated, this experiment is designed to collect audio and written responses from the participants by describing what happened next to the main character of the demonstrated short stories. The aim is to collect spoken and written narratives where both types are produced in a spontaneous and informal way by using the same mode as the stimulus.

We create three crowdsourcing tasks, each containing 20 stimuli, where 10 of them are presented in audio and the other 10 in written format. Either six or seven of the stimuli in each crowdsourcing task are filler stories, both in audio and written forms. All three crowdsourcing tasks are available to every participant. We ask them to complete all three tasks (i.e., 60 stimuli) (see the instruction in 5.22). Some participants complete all three tasks, but we also receive responses for only one or two crowdsourcing tasks from some participants, which we also consider in the analysis.

(5.22) Welcome to our **story continuation** micro task series.

> In this task (out of 3 tasks in total), you will see 20 very short story beginnings of 1 to 8 sentences length. You are to conceive a continuation of these stories, with at least 1 and at most 3 sentences. We will present the stories in different ways to you. You are to respond to written stories by writing, and to the audible stories by short speech recordings.
>
> For every story, please try to **imagine what happens next to the main character**, and follow what comes to mind quickly, without any elaborate planning! This task will last about 15 minutes. Afterwards, you can immediately work at the remaining 2 more tasks which are all available in your job list, so best if you block approx. 45mins of your time for the whole task series right now.

The interface of the main experiment is similar to the pilot study described in Section 5.2.2. We use the same instructions as in 5.10 and 5.11 at top of the pages where we present the written and audio stories, respectively.

The experiment is executed in October-November 2020 on two mutually exclusive groups of participants (Group1 and Group2). Both groups are exposed to the same 60 stimuli yet in different modes. For instance, if the story "Patricia rode all day yesterday" is shown to the participants of Group1 as a written stimulus, Group2 is exposed to its audio version. Consequently, we do not collect narratives for the same spoken and written stories from the same individuals, which was a necessary constraint for getting spontaneous responses at each step. We present the stimuli in a random order but distribute the filler stories uniformly between the stories, so that the participants are not exposed to the actual stimuli repeatedly.

Table 5.2 summarizes the participation statistics for the experiment. Group1 is composed of 131 and Group2 is composed of 114 participants. 220 crowdsourcing tasks (i.e., 20 stimuli in each) were submitted by Group1, whereas 239 task submissions were received from Group2. We excluded the responses to filler stories and processed 3054 responses for spoken stories and 3049[12] for written stories, in the analysis.

As mentioned earlier, we provide 3 crowdsourcing tasks (i.e., 60 stimuli) for all the participants. Although we asked them to complete all the tasks, we do not make it mandatory as the tasks can take a considerable amount of time. As a result, we received a varying number of task submissions from the participants, as presented in Table 5.2. For instance, 52 participants in Group1 and 34 participants in Group2 submitted only one crowdsourcing task (i.e., 20 stimuli). The number of participants who completed all three tasks is unfortunately very small in Group1 compared to Group2 (10 vs 45). During the execution of the experiment for Group1, a technical issue occurred with showing the third task to the participants. Consequently, many of the participants in Group1 could not see

[12] There exist empty responses in the written cases, this is why there is a difference between the number of responses for spoken and written modes.

		Group1	Group2	Total
1	Participants	131	114	245
2	Submitted crowdsourcing tasks	220	239	459
3	Number of spoken responses (including responses to fillers)	2200	2390	4590
4	Number of written responses (including responses to fillers)	2200	2390	4590
5	Number of spoken responses (excluding responses to fillers)	1465	1589	3054
6	Number of written responses (excluding responses to fillers)	1465	1584	3049
7	Participants completed 3 tasks	10	45	55
8	Participants completed 2 tasks	69	35	104
9	Participants completed 1 task	52	34	86
10	Minimum number of responses for a stimulus	70	73	70
11	Maximum number of responses for a stimulus	76	85	85

Table 5.2: Main experiment participation summary

nor complete the third task. However, we resolved the issue by increasing the number of participants in Group1 to keep at least 70 responses for each story in our stimuli set, which is demonstrated at row 10 in Table 5.2.

At the end of each crowdsourcing task, we provide a free text area for the participants to express their feedback about the task. We receive feedback from 28 participants. Two of the comments are about a temporary technical problem in the interface they encountered, one was complaining about the amount of payment. The rest of the comments are all very positive about the experiment (e.g., “I really enjoy doing these tasks!”, “Another good study which I enjoyed doing.”, “This was fun! :-)”). With this positive feedback and the observed coherent nature of the responses, we are confident that our participants experienced gratifying motivation and engagement for the experiment. The feedback also indicates that in addition to the payment they received, some of the participants were also motivated by finding entertainment in our experiment.

5.2.4 Transcription

We follow a semi-automated procedure to transcribe the audio responses. First, the audio recordings were automatically transcribed via Google speech-to-text API.[13] We then go over all the transcriptions manually, listen to the recordings one by one and correct the errors in the auto-transcriptions. We also add manual punctuation marks (sentence final markers and commas) when needed into the transcribed audio responses.

In 18% of the cases, auto-transcription is completely correct as in 5.23. Therefore, we do not make any changes to them.

(5.23) Jack was excited about his trip and went to go and get a cup of coffee while waiting for his train.

In 9% of the cases, the transcription is correct but we need to add some punctuation and capitalize the sentence initial letters, for better structuring. For instance, the transcription in 5.24 was modified as in 5.25.

[13] `https://cloud.google.com/speech-to-text`

(5.24) but the samples might have been contaminated just like last time this time John would have to be more careful.

(5.25) But the samples might have been contaminated just like last time. This time John would have to be more careful.

In the remaining 73% of auto-transcriptions, we make corrections in the transcribed text. Some of the transcriptions include erroneous wording as demonstrated in 5.26, which is corrected as in 5.27, and some others miss speech pieces as demonstrated in 5.28, which is corrected as in 5.29.

(5.26) Jon Stewart the samples so he could take a better look down at the station.

(5.27) John stored the samples so he could take a better look down at the station.

(5.28) don't stash is at your friends over there to the right look at that so guy curly hair looks like Colin Firth.

(5.29) Don't stare, she said to her friends but over there to the right look at that tall guy, curly hair. He looks like Colin Firth. Do you think it is him?

Non-verbal communication like laughter and pause, and prosodic features such as stress intonation, sound prolongation, emphasis, etc. are not marked in the transcriptions. We eliminate 5% of the responses after listening the audio responses. Most of these responses provide silent or noisy audio recordings and the others generate texts on topics different from the stimulus.

5.2.5 Analysis

For the analysis, we process the responses submitted by the participants as story continuations. Table 5.3 shows the statistics on the length of responses we receive. The lengths of spoken responses are computed after we inserted the final punctuation marks into the transcribed texts. As shown in rows 1,2, the responses to written stories are longer on average than responses to spoken stories. Responses tend to be longer as the length of the stimuli increases. Especially the difference in response length for 1-clause stimuli and 5- and 8-clause stimuli is remarkable.

In the rest of this section, we present more statistics to better describe the responses and the way we process this data.

5.2.5.1 Syntactic Properties of Responses

We automatically segment the responses into sentences using the nltk library [Bird et al., 2009] in Python. We then run the Berkeley Neural Parser [Kitaev and Klein, 2018] on our data and extract clause and noun phrase (NP) structures using the constituency parse trees generated by the parser. We do not do any manual correction on the parse trees although we observe considerable number of errors in them. Thus, the numbers presented in this section should be considered as a rough approximation of data description rather than a robust analysis.

Table 5.4 gives quantitative information on the syntactic structures in the responses. Subordinate clauses are more common for both modes but the difference between the number of clause types is more prominent for the written mode. In addition, written responses contain less, and therefore longer, clauses than the spoken mode, given that the written responses are longer than spoken responses. As subordination is more common for

		Spoken	Written	Total
1	Average response length (chars)	120	133	126
2	Average response length (words)	23	25	24
3	Average response length for 1-clause length stimuli (words)	19	23	21
4	Average response length for 2-clause length stimuli (words)	23	24	23
5	Average response length for 3-clause length stimuli (words)	23	25	24
6	Average response length for 5-clause length stimuli (words)	24	25	25
7	Average response length for 8-clause length stimuli (words)	24	25	25

Table 5.3: Main experiment response length statistics

both modes, this result is not in line with what has been reported earlier, for instance, by Schäpers [2009]. However subordinate clauses are more dominant and longer in written responses, which is compatible with the results reported in Schäpers [2009].

Third person pronouns and non-pronominal NPs have similar frequency distributions in both modes, with third person pronouns being more common in the spoken mode than in the written mode with the slight difference of 1%. The length of the non-pronominal NPs (in terms of words) is computed as 2.89 for the spoken and 2.86 for the written mode. This result does not confirm our findings in Chapter 4 where we report that NPs are longer in written mode. However, because the differences in these comparisons are not substantial enough to eliminate the inaccuracies caused by automatic parsing, we avoid making generalizations from them.

		Spoken	Written	Total
1	Coordinate clauses	3261	2127	5388
2	Subordinate clauses	3787	3652	7439
3	Third Person Pronouns	7266	7828	15094
4	Non-pronominal NPs	17592	19999	37591
5	Percentage of Third Person Pronouns[14]	29%	28%	29%
6	Percentage of Non-pronominal NPs	71%	72%	71%
7	Non-pronominal NP-length	2.89	2.86	2.87

Table 5.4: Syntactic Properties

5.2.5.2 Discourse Features in Responses

In order to explore the type of responses and their discourse relation to the stimuli, we created a subset of the data containing approximately 9% of the responses (i.e., 500 in total) with a balanced number of spoken and written responses. Also, the sample is balanced for continuation length. We marked the responses in this subset using a scheme inspired from the Temporal and Expansion discourse relations (e.g., in the Penn Discourse TreeBank framework, [Webber et al., 2018]). We define three distinct groups marked by

E1, E2 and T as described below (S and R denote the clauses in the stimuli and the responses, respectively):

1. [S1 .. Sn] TEMPORAL [R1 .. Rn] (T): There is one event in S, and a temporally-later event in R. The stimulus event is usually stated in the first clause S1, while S2..Sn are additional elaborations. So, the relation holds between the complete stimulus and the response, as illustrated in Example 5.30.

   (5.30) **Stimulus**: "Matthew went out for the first time after the lockdown was lifted. Restaurants were open and busy. The streets were empty because it was very cold outside."
   Response: "He entered the warm restaurant to be met by the smell of good wine and strong garlic. He breathed in the scent, relieved as a feeling of normally settled over him."

2. S1 .. [Sn] EXPANSION [R1 .. Rn] (E1): R adds some elaboration to the final clause of S. Importantly, the response does not introduce a new event (i.e., the relation is not TEMPORAL), as illustrated in Example 5.31.

   (5.31) **Stimulus**: "Carol made a peppermint tea before the sun rose. Herbs can balance the emotions. Peppermint helps for anxiety so it was a perfect choice for this morning."
   Response: "Peppermint also helps with bloating, and she suffers with this from IBS."

3. [S1] .. Sn EXPANSION [R1 .. Rn] (E2): R adds some elaboration to the first clause in S. So, the relation holds between the first stimulus clause and the response. Importantly, R does not introduce a new event (i.e., the relation is not TEMPORAL), , as illustrated in Example 5.32.

   (5.32) **Stimulus**: "Theresa published an article in Lancet. Lancet is a leading medical magazine. It covers everything about health. It publishes up-to-date studies. There is a review process before the articles are issued. Lancet accepts only original papers. Therefore, this is an impressive success."
   Response: "Her article was about covid-19. She was very excited to have this published in Lancet."

As shown in Table 5.5, the most common relation in our subset is TEMPORAL (T). In the written responses, we observe a higher percentage of temporal relations than in the spoken responses.[15]

	Relation Type	Spoken (%)	Written(%)	Total(%)
1	E1	12	10	11
2	E2	27	17	22
3	T	61	73	67

Table 5.5: Discourse relations between the responses and the stimuli

Distribution of referential forms for the first reference to the main character in these responses is presented in Table 5.6. The statistics show that E1 and T type of relations trigger more frequent use of full NPs (names in this case) than the E2 relations.

		E1	E2	T
1	Name	36	14	27
2	Pronoun	64	86	73

Table 5.6: Referential choice with respect to discourse relation type

We fit a generalized mixed-effects logistic regression [Bates et al., 2015] to this subset of data to predict the binary referential choice (Pronoun vs Name) from the relation type, with random effects for the participants and the stimuli. For the relation type, sum coding is used, with E1 being coded as (1,0), E2 being coded as (0,1), and T being coded as (-1,-1). The model results are reported in Table 5.7. The model indicates there is a correlation between the relation type and the referential form. For instance, pronouns in E2 relations are used more frequently than the average pronoun usage for all the relation types ($p<0.05$) which is in line with the frequency values presented in Table 5.6.

	Estimate(log-odds)	Odds-ratio	S.E.	z val.	p
Intercept	1.27	3.53	0.22	5.76	***
relation-type1	-0.48	0.62	0.29	-1.64	0.10
relation-type2	0.56	1.75	0.26	2.12	*

Table 5.7: Summary of the generalized mixed-effects logistics regression model (Significance thresholds: ***$p<0.001$, **$p<0.01$, *$p<0.05$).

5.2.5.3 Processing of Responses

For the investigation of our main research question, we mark the first references to the main character in the responses. For instance, first references to the protagonist in stimulus 5.33 (i.e., Robert) are marked as in responses 5.34 and 5.35.

(5.33) Robert finally solved the mystery.

(5.34) ROBERT gasped as he realised the truth. All this time his boss had told him to stop investigating and now Robert knew why.

(5.35) It suddenly came to HIM out of nowhere as he looked out of the train window watching the fields, and forests go by.

We do not count the pronouns used to refer to the main character but which do not match the gender. For instance, 5.37 is submitted as a continuation for 5.36 and the possessive pronoun "his" is aimed to refer to "Mary" but it does not match the prototypical gender indicated by the name "Mary". Therefore, we do not include this response in our analysis. We received 36 responses with an unmatching gender reference.

(5.36) Mary received good news today!

[15] In Section 5.4, we provide a brief discussion on this difference by referring the work of Klein and Stutterheim [1992].

(5.37) his mother came to visit him at the hospital and told him the doctor was going to release him today.

For some names, participants occasionally use short or casual forms. For instance, “Fred” or “Freddie” were used instead of “Frederick” and “Danny” instead of “Daniel”. We count these cases as instances of reference by the name.

If there exist no reference to the main character in the response, as demonstrated in 5.39 which is a continuation to the story 5.38, we do not consider that responses in the analysis. We received 443 responses without a reference to the main character.

(5.38) Tony started organic farming for health reasons. Organic food is healthy because no chemicals are used in the process. It yields more food with less expense when it is done systematically. The products are much tastier: Tomatoes smell wonderful, peppers are flavorful.

(5.39) Everything tastes fresh and natural. Things grow as it should be grown. Nature has a way of nurturing her harvest.

5.3 Results

5.3.1 Descriptive Analysis

As mentioned earlier, we only use proper names to introduce the main characters and do not provide additional details except for implying the person’s gender by using common male and female names. The participants used either a personal pronoun or the name for referring to the main character. We do not encounter any other referential forms denoting the main character in the responses, such as definite or indefinite noun phrases.

We compute the distribution of the form of referential expression (pronoun vs explicit name) of the first reference to the main character in the responses. The percentage of pronouns as the first reference (pronoun usage, henceforth) is 79.6% in spoken and 77.1% in written mode. The average story length for which the participants used a pronoun in the continuations as the first reference to the main character is 3.56 clauses for the spoken and 3.44 clauses for the written mode.

Distribution of pronoun usage with respect to the stimuli length (in terms of clauses) is shown in Figure 5.1. It demonstrates that pronoun usage decreases with the increasing length of the stimuli for both modes. The chart also indicates that pronoun usage is higher in spoken responses than in written responses for longer stories.

In order to get more information about the internal patterns of responses, we further show the distribution of pronoun usage for each story length via box plots (Figure 5.2) and bar plots demonstrating pronoun usage for each stimulus (Figures 5.3-5.7).

Figure 5.2 indicates that the median values of pronoun usage are greater in the spoken data for 2-,3-,5- and 8-clause-length stimuli, whereas the median is obviously higher in written data for 1-clause-length stories. The median of the written responses for 1-clause-length is also greater than all other values in the corresponding spoken responses. For the 8-clause-length stories, the spoken median is only slightly smaller than the upper quartile of the written data and the upper quartile of the spoken data is greater than all the values in the distribution of written responses. As a result, we conclude from this box plot representation that pronoun usage is higher in written responses for very short 1-clause-length stories but higher in spoken data for long 8-clause-length stories. We cannot make such clear inferences from the box plots demonstrating the distributions for 2-, 3-, and 5-clause-length stories.

Figures 5.3-5.7 support our interpretation of box-plots in Figure 5.2. We observe that pronoun usage in written responses for all the stimuli is higher than in spoken responses

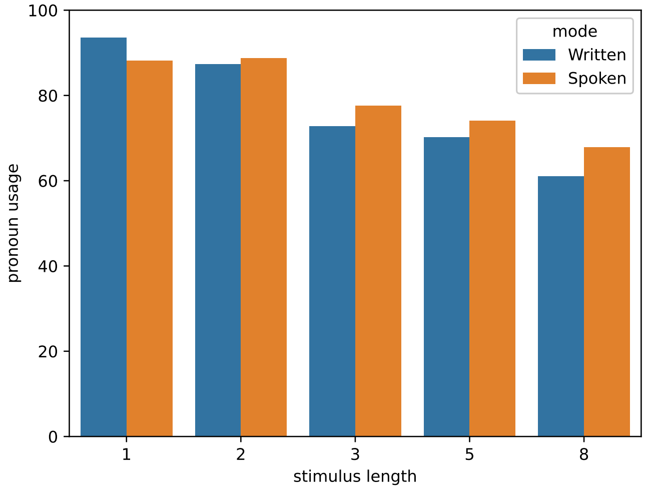


Figure 5.1: Distribution of pronoun usage for each stimulus length

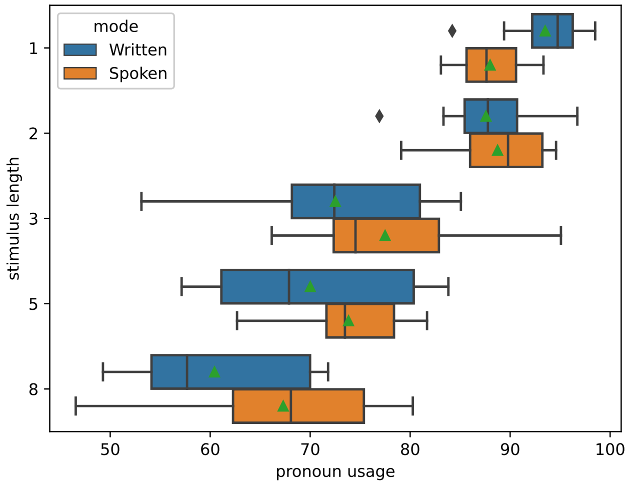


Figure 5.2: Box-plots of pronoun usage for each stimulus length

for 1-clause-length stories, whereas the responses to 8-clause-length stories exhibit the opposite behavior; pronoun usage is higher in spoken responses for all the stories except one (S34) for the 8-clause-length condition. We do not observe such clear patterns for 2-, 3-, and 5-clause-length stimuli, which are shown in Figures 5.4-5.6.

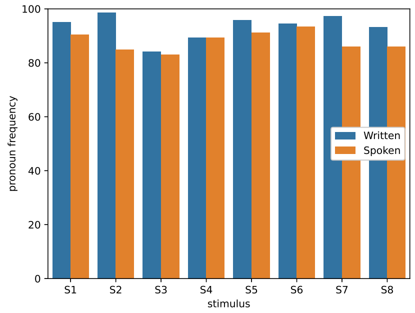


Figure 5.3: Pronoun usage in responses to 1-clause length stimuli

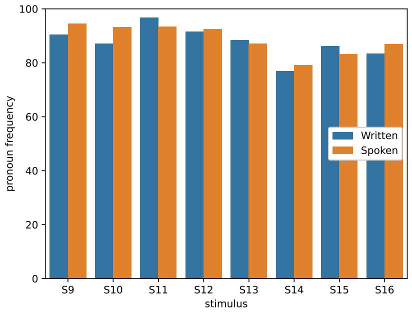


Figure 5.4: Pronoun usage in responses to 2-clause length stimuli

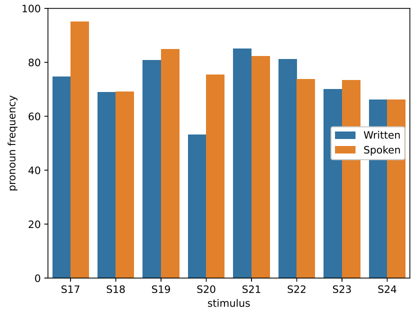


Figure 5.5: Pronoun usage in responses to 3-clause length stimuli

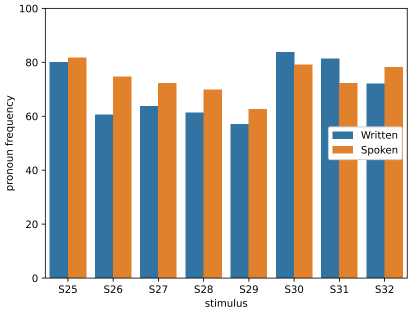


Figure 5.6: Pronoun usage in responses to 5-clause length stimuli

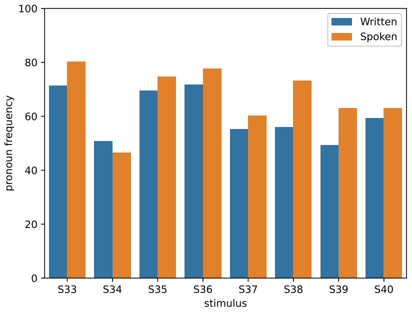


Figure 5.7: Pronoun usage in responses to 8-clause length stimuli

5.3.2 Generalized Mixed-Effects Logistic Regression

We fit a generalized mixed-effects logistic regression model to the responses in order to evaluate the factors influencing the referential choice, using the lme4 package [Bates et al., 2015] in R. The dependent variable (i.e., the first referential choice for the main character in each response) is a binary variable, with the possible values of Pronoun and Name. We considered the “stimulus length” as a numerical variable rather than a categorical one and centered it around its mean. We applied sum coding for the “mode”, with the “spoken” mode being coded as 1 and the “written” mode being coded as -1.

We increased the complexity of the model gradually. Table 5.8 shows the summary and comparison of the models via likelihood ratio tests (done by the anova() method in R). We

model	formula	AIC	$R^2(f)$	$R^2(t)$	Anova	p
m0	ref_expr∼mode*length	5554	0.05	-	-	-
m1	ref_expr∼mode*length+ (1|participant)	4518	0.08	0.51	m1,m0	***
m2	ref_expr∼mode*length+ (1|participant)+ (1|stimulus)	4384	0.09	0.56	m2,m1	***
m3	ref_expr∼mode*length+ (1+mode|participant)+ (1|stimulus)	4381	0.09	0.57	m3,m2	*

Table 5.8: Comparison of the generalized mixed effect models fitted to the data (Significance thresholds: ***p<0.001, **p<0.01, *p<0.05) The column $R^2(f)$ shows McFadden's pseudo-R^2 value for fixed effects, whereas the $R^2(t)$ column represents the conditional R^2, which describes the proportion of variance explained by both the fixed and random effects. The R^2 numbers in the table indicate that the model without random factors fits to our data very poorly (i.e., explanatory power is 5% for m0). In contrast, the total explanatory power of a mixed-effects model is substantial (conditional R^2 is 57%) and the part related to the fixed effects alone ($R^2(f) = 9\%$) is also higher than the $R^2(f)$ of the fixed-effects logistic regression (m0).

first fitted a fixed-effects model (m0). Mode, length of the stimuli and their interactions are considered as fixed effects in this model. Next, in order to address the differences between participants, we added a random intercept of participants to the model (m1). The likelihood ratio test indicates that m1 is a better model because it has a smaller AIC value, as shown in Table 5.8. In the next step, we added a second random effect for the stimuli (m2) which also improved the model significantly (p<0.001). Adding a random slope over "stimuli" leads to the singularity of the model. A random slope for "mode" over the "participants" (m3) does not cause singularity and contributes to the model significantly (p<0.05). Therefore, we include random slope for the mode over participants. We did not include the participants' features in the model such as gender and age because we do not control these variables in our experiment.

The summary for the fixed effects of the final model (m3) is demonstrated in Table 5.9. The Intercept in Table 5.9 indicates that the mean value for pronoun usage across the dataset is substantially greater than name usage in the fitted model for the average length of stimuli (odd-ratio=6.87). Participants seem to use fewer pronouns in written mode than in spoken mode for the average length of stimuli, and the difference between modes is statistically significant. The model coefficients indicate that an increase in the length of the stimuli decreases the pronoun usage significantly (p<0.001) for both modes. The decrease in pronoun usage, as demonstrated in Figure 5.8, is more prominent for the written mode than the spoken mode for the increasing length of stimuli (p<0.05). This inference is compatible with the data demonstrated in Figure 5.2, which indicates that pronoun usage in the written mode is affected more than the spoken mode by the increasing length of the stimuli.

The random intercepts for the participants are demonstrated in Figure 5.9. We removed the labels for participant ids for the sake of simplicity. The random intercepts for the participants demonstrate a figure with a longer tail on the left than on the right, which indicates that there are outlier participants who tend to use much less pronouns than the

	Estimate(log-odds)	Odds-ratio	S.E.	z val.	p
Intercept	1.93	6.87	0.16	12.05	***
mode	0.09	1.10	0.092	1.97	*
length	-0.32	0.73	0.04	-7.53	***
mode:length	0.03	1.03	0.036	2.05	*

Table 5.9: Summary of the final model (Significance thresholds: ***p<0.001, **p<0.01, *p<0.05).

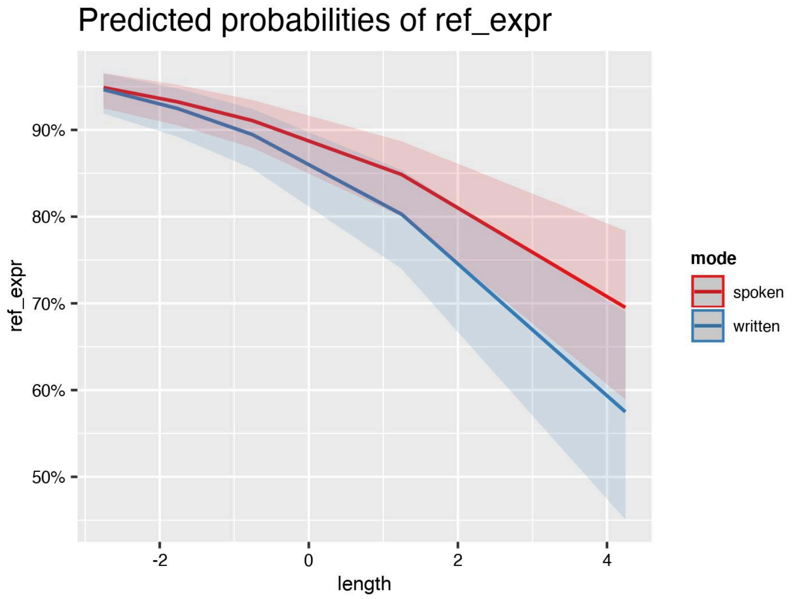


Figure 5.8: Predicted probabilities of pronoun usage with increasing length (centered values) across modes.

others. Excluding these outlier participants in the computational and statistical analyses might give more insight about the impact of other features. Due of time constraints, we decide to leave it for a future task.

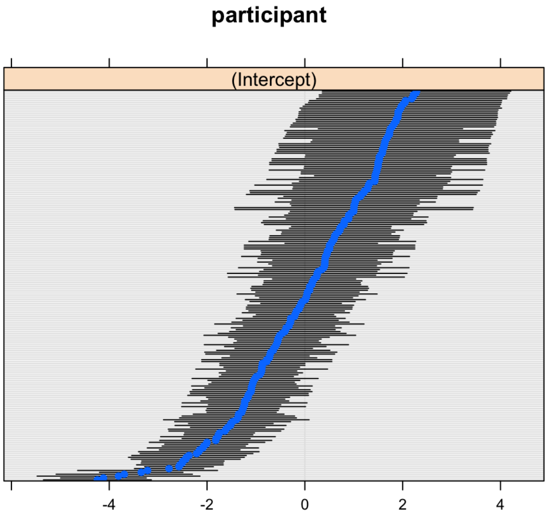


Figure 5.9: Random intercepts with the confidence intervals for the participants

5.3.3 Classification Experiments

Same and van Deemter [2020] perform an extensive evaluation of the feature sets used in a variety of computational systems for identifying referring expressions. Their evaluation reveals the importance of four feature sets, three of which are related to antecedents (animacy and plurality, grammatical role, and form), which we control and keep constant in our stimulus design. The fourth prominent feature set that Same and van Deemter [2020] use is recency, which is the varying factor in our experiment. We believe that computational classification experiments run on our data can give insights about what other aspects will be relevant in addition to recency. This is a particularly interesting question given that other potentially important features are kept constant. We run preliminary experiments on our data for the task of predicting the form of the first reference to the main character (first reference, henceforth in this section) and report the initial results in this section.

From the dataset, we extract 36 features that can be grouped into five categories: features related to the stimuli, the protagonists in the stimuli, the participants' responses, the first references, and the participants. Some of these features are numerical, such as the length of the stimuli or the age of the participant. In addition to these numerical features, we also generate categorical variables which convert the numerical data to a categorical form. For instance, the participants' ages are categorized into 8 groups, which includes the number of individuals younger than 20, between 20 and 25, 25 and 30, 30 and 35, 35 and 40, 40 and 45, 45 and 50, and older than 50. Together with the categorical representations of numerical features, we use in our classification experiments a total of 41 features, which are provided in Table 5.11.

Rows 1-7 in Table 5.11 relate to the stimuli, the text and the length of the text of the stories as well as the mode that the stimuli is provided in (i.e., spoken or written). Rows 8-10 encode the protagonist's name and the gender implied by the name. Rows 11-26 are related to the responses submitted by the participants, including features encoding the length of the response and the context in which the first reference occurred. The context is provided by considering the unigram, bigram, trigram tokens before and after the reference, as well as parts of speech tags of those tokens. Rows 27-35 are related to the first reference, including syntactic features (e.g., the dependency relation established by the root node in the parse tree and whether the reference is subject or object) and the token order of the first reference in the text. Rows 36-41 are related to the participants. We represent each participant with a unique id. Additionally, characteristics of the participants (i.e., gender, age, education level, and spoken English variety) gathered during the experiment's execution are also used in the classification experiments. For extracting the features concerning the syntax (i.e., grammatical role or part of speech tag), we use the automatic PoS identification and parsing provided by a Python NLP package, **stanza** [Qi et al., 2020].

We explore to what extent we can predict the form of the first reference (i.e., pronoun or name) in each response by using a machine learning-based classifier. We randomly divide the data into training and test sets, where the training set covers 80% and the test set covers the remaining 20%. 77.7% of the test set contain the first references in the form of a pronoun. Therefore, the majority baseline is 77.7%, as is the case when the same label is assigned to all samples without making a real classification. We then experiment with a number different classifiers provided by the **sklearn** package in Python [Pedregosa et al., 2011], including RandomForestClassifier[16], AdaBoostClassifier[17], and GradientBoosting-

[16] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

[17] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html`

Classifier[18]. We start the experiments using all features and then gradually prune the least important ones in the set of features according to the initial fitted models.[19]. We obtain the best result with the GradientBoostingClassifier[20] using only 3 of the features. The results are shown in Table 5.10. The results in Table 5.10 indicate that the classifier that employs all features provides a slight performance boost (i.e., 0.5%) over the baseline. However, employing only a subset of the three features, namely, length of the stimuli, PoS tag of the token that follows the first reference in the response and the participant identifier, improves the performance significantly (by 6.9%). Relative importance scores of these features when making a prediction is found as follows: story_length_clause:0.18, response_ngram1_post_pos: 0.22, participant_id: 0.60.

The initial experiments we report in this section indicate that using a subset of the features may be more beneficial than employing all extracted features. One reason for this result is the fact that using many features, in some cases, causes the problem of overfitting on the training set. In addition, the outcome of the automatic parse results, and hence, the syntactic information we use in the experiments are erroneous. Inconsistent labels, in our opinion, may lead to confusion for the classifier. Correcting the parse trees could potentially provide better results in this classification task. For further exploration, we believe that grouping PoS tags with a similar function into single categories might be beneficial. Additionally, since the participant's identity emerges as a key factor in our classification experiments, we believe it would be intriguing to investigate the scenario in which the outlier participants mentioned in the preceding section are excluded from the data. We reserve analysis of these experimental settings for further investigations.

	Experiment	Features	Score
1	Baseline	-	77.7%
2	GradientBoostingClassifier	All Features	78.2%
3	GradientBoostingClassifier	3 features (story_length_clause, response_ngram1_post_pos, participant_id)	84.6%

Table 5.10: Classification experiment results

[18] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html`

[19] Checked using the feature_importances_ attribute of the fitted model

[20] The best configuration obtained with these settings: number of estimators=400, learning rate=0.2, max depth = 2, random state=1

	Feature Category	Feature	Explanation
1	Stimulus	mode	Production mode (audio or written)
2		story_content	Text of the stimulus
3		story_length_word	Length of the stimulus (words)
4		story_length_word_cat	Categorical repr. of item 3
5		story_length_clause	Length of the stimulus (clauses)
6		story_length_sentence	Length of the stimulus (sentences)
7		story_length_markables	Number of markables in stimulus
8	Protagonist	protagonist_name	Name of the protagonist
9		protagonist_name_length	Length of the name (chars)
10		protagonist_gender	Typical gender implied by name
11	Response	response_length_char	Length of the response (chars)
12		response_length_char_cat	Categorical repr. of item 11
13		response_length_word	Length of the response (words)
14		response_length_word_cat	Categorical repr. of item 13
15		response_ngram1_pre	Unigram text before the reference
16		response_ngram1_post	Unigram text after the reference
17		response_ngram2_pre	Bigram text before the reference
18		response_ngram2_post	Bigram text after the reference
19		response_ngram3_pre	Trigram text before the reference
20		response_ngram3_post	Trigram text after the reference
21		response_ngram1_pre_pos	Unigram PoS tags before the reference
22		response_ngram1_post_pos	Unigram PoS tags after the reference
23		response_ngram2_pre_pos	Bigram PoS tags before the reference
24		response_ngram2_post_pos	Bigram PoS tags after the reference
25		response_ngram3_pre_pos	Trigram PoS tags before the reference
26		response_ngram3_post_pos	Trigram PoS tags after the reference
27	Reference	reference_deprel	Dependecy relation to the head
28		reference_token_order	Token order in the response
29		reference_token_order_cat	Categorical repr. of item 28
30		reference_is_first_token	Is the reference first token in response?
31		reference_is_second_token	Is the reference second token?
32		reference_is_third_token	Is the reference third token?
33		reference_is_subj	Is the reference a subject?
34		reference_is_obj	Is the reference an object?
35		reference_is_sentence_initial	Is the reference first token in the sentence?
36	Participant	participant_id	Unique identifier for participant
37		participant_gender	Gender of the participant
38		participant_age	Age of the participant
39		participant_age_cat	Categorical repr. of item 39
40		participant_degree	Education level of the participant
41		participant_english	English variety the participant speaks

Table 5.11: Features used in the classification experiments

5.4 Discussion

Our first observation on our analysis concerns the type of texts produced in response to stimuli. We presented an analysis of discourse relations on a subset of the experiment data in Section 5.2.5.2. According to the findings, 67% of all responses introduce a TEMPORAL relationship to the stimuli by introducing a temporarily-later event. According to the framework introduced by Labov and Waletzky [1967], these responses form a temporal juncture with the event introduced in the stimulus, and therefore these stimulus-response pairs can be considered as narratives (i.e., Labov and Waletzky define narrative as a sequence of clauses which contains at least one temporal juncture). Our analysis also shows that written responses have more TEMPORAL discourse relations between stimuli and responses than written responses. The interplay between the realms of narrative structure and reference is studied by Klein and Stutterheim [1992], who use the term "referential movement" in stories consisting of a "main structure" that is at various points interrupted by portions of "minor structure" (descriptive, elaborative content where the temporal references do not move forward in time). We see our distinction between coherence relation types as making the same point: In the narrations, the "T" cases represent a switch from a preceding "minor structure" back to the "main structure". Klein and Stutterheim [1992] to our knowledge did not investigate the role of pronominalization, but we assume that the use of a full NP as opposed to a pronoun serves as a signpost to the audience for making the switch back to the main structure. Constructing this signpost, we furthermore assume, requires planning work that is undertaken more often in writing than in speech, which is compatible with the observation that T discourse relations are established more often in written responses than spoken responses in our experiment.

The second observation we made based on the results shown in Section 5.3 is regarding the dominant use of pronouns for both modes as a first reference to the main character in responses. The statistics reported in Section 5.2.5.1 suggests a pattern similar to what is reported in Chapter 4, which indicates that non-pronominal NPs are more common than third person pronouns for both modes. However, the expression used for the first reference to the main character in our responses does not reveal a similar pattern because at least 60% of the first references are pronouns for both modes. This result is not directly comparable to the results we report in Chapter 4 (i.e., this result only applies to the first referential form referring to the main character in the responses, not to all referring expressions), but its significant departure from the statistics in Section 5.2.5.1 is remarkable. Our findings are comparable to and consistent with the findings of Fox [1987], who reveals that in the absence of interfering referents between the antecedent and the anaphoric expression, pronoun usage is significantly prevalent in both modes. [Arnold and Griffin, 2007, p.521] state that "even when a pronoun would not be ambiguous, the presence of another character in the discourse decreased pronoun use". We believe the high number of pronouns referring to the main character for both modes (i.e., no difference between spoken and written) might result from the fact that we introduce only one human (or animate) entity in our stimuli, so there is no competition with another entity.

The third observation concerns the relative distribution of all third person pronouns and non-pronominal noun phrases across modes. Frequency distributions regarding all NPs in the dataset presented in Section 5.2.5.1 indicate that the use of third person pronouns and non-pronominal NPs are similar for both modes, with pronouns being slightly more frequent in spoken than in written responses (i.e., 29% vs 28%). The frequency distribution of the same elements when used as a first reference to the main character shows a similar trend in that pronouns are more common in spoken than in written responses (i.e., 79.6% vs 77.1%). The statistical model we fitted to our data confirms that spoken mode contains more frequent use of pronouns, and this difference is significant ($p<0.05$) for the average length of the stimuli. As a result, our findings on the pronoun usage as a first reference to the main character indicates that third person pronouns are more frequent in the spoken mode than in the written mode on average. For the result concerning all NPs, the numbers are very similar (29% vs 28%) and we obtain these results through automatic parsing, which contains errors by nature; we therefore interpreted the NP distribution difference between modes as inconclusive.[21]

The fourth observation is regarding the distribution of pronoun usage for varying lengths of the stimuli. Figures 5.3-5.7 show that the patterns for 1- and 8-clause length stimuli are regular, with written responses containing more pronoun usage than spoken responses for 1-clause stimuli and the opposite tendency for 8-clause stimuli. A closer look at Figures 5.3-5.7 reveals irregularities in the seemingly regular patterns. For instance, written responses to stimulus S34 in Figure 5.7 (given in 5.40) include more pronoun usage (51%) than the spoken responses (47%). Consequently, S34 does not fit the general pattern of 8-clause-length stories. In addition, this stimulus triggers less pronouns than the other stimuli with the same length. Another stimulus (S36) in Figure 5.7, which is given in 5.41 represents a regular pattern with much higher pronoun usage.

(5.40) Danny moved to an island. It was rather small. Vehicles were not permitted there. It was a mysterious place, too. There was a volcano in the South. One night, the volcano became active, then the sky brightened. That was really scary.

(5.41) Gabriela moved to a new neighborhood. The previous area was chaotic because it was very central. This new neighborhood is quiet. There are parks nearby. The new house is great. It is quite peaceful. A whole new life can start here.

In our experimental procedure, the event brought about by the main character is introduced in the first clause for the stimuli longer than one clause, which is usually followed by elaborative clause(s) establishing EXPANSION relations (see Section 5.2.5.2) with the first clause. Different from the general pattern, the stimulus in 5.40 contains a second event temporarily ordered (a “complicating action” clause according to Labov and Waletzky [1967]) with respect to the main character’s action introduced in the first clause. The more regular stimulus in 5.41 does not contain a complicating action, and therefore cannot be considered as a narrative according to Labov and Waletzky [1967]. The subset we used for the discourse relation evaluation in Section 5.2.5.2 contains 18 responses to stimulus 5.40 and 14 responses to stimulus 5.41. Responses to 5.40 contain only E1 (11%) and T (89%) type of relations, whereas responses to 5.41 contain all three relation types (14%, 14% and 72% for E1, E2 and T, respectively). This indicates that responses to 5.40 does not elaborate on the main character’s action introduced in the first clause, but introduces new events or expand the second action in the stimulus, whereas responses to 5.41 relate to the stimulus with expansion relation (E2) elaborating the main action (“orientation” according to the framework of Labov and Waletzky [1967]). This observation indicates that the type of the stimuli (e.g., narrative or not) has an impact on the type of relation established by the response. Therefore, considering the correlation between the relation type and referential choice discussed in Section 5.2.5.2, we can hypothesize that the type of the stimulus affects the form of the first reference to the main character. However, this hypothesis needs further investigation, with all the stimuli being analyzed for their qualitative features and the relation types with their responses being labelled for a larger number of instances.

The fifth observation on the data presented in Section 5.3 is that pronoun usage decreases with increasing anaphoric distance for both modes, which is consistent with

[21] For a more thorough study of the data, we believe using the manually corrected parse trees in the analysis could be beneficial, but we’ll leave that for a later endeavor.

previous comprehension research [Clark and Sengul, 1979, Streb et al., 2004], with the assumption that longer pronoun comprehension time means less pronoun usage in production, for a setting with long anaphoric distance. The difference in pronoun usage for the written mode between 1-clause-length and 8-clause-length stories (in Figure 5.1) is 32%, whereas it is 20% for the spoken mode. This finding is consistent with prior corpus-based research (e.g., Fox [1987]) and our findings in Chapter 4, which indicates that spoken texts allow for greater anaphoric distance than written texts. In our interpretation, the difference in the decrease of pronoun usage between spoken and written modes indicates that anaphoric distance has a larger effect in referential choice for the written mode, which is also confirmed with the generalized mixed-effects analysis in Section 5.3.2.

5.5 Summary and Conclusion

The aim of this study is to investigate the differences across spoken and written modes, in terms of the impact of anaphoric distance on the referential choice in language production. Relevant contrastive corpus-based studies (e.g., Fox [1987]) indicate that spoken texts exhibit longer average anaphoric distance than written texts between third person pronouns and their antecedents, where the distance is measured in terms of clauses. In relation to these findings, we hypothesize that spoken texts contain more pronouns than written texts in the case of long anaphoric distances. We design an experimental setup to test this hypothesis, where we manipulate the anaphoric clause-based distance systematically while keeping constant the features of the antecedent, such as animacy, plurality, grammatical role and the form of referential expression (i.e., name or pronoun). Previous studies (e.g., Schäpers [2009]) demonstrate that clause lengths, types and tokens differ between spoken and written texts. We manipulate the distance in terms of number of clauses, but control the length of clauses, type of tokens and the coherence relations linking the clauses, in order to understand the impact of the mode.

To the best of our knowledge, this is the first experimental study investigating the mode differences with respect to the impact of anaphoric distance in a production task.

We use short stories varying in length as the stimuli and expect the participants to continue these short stories by imagining “what happened next to the main character?”. This experiment has been conducted via crowdsourcing. Only 5% of the responses have been eliminated due to, for instance, bad audio quality or the topic differing from the stimulus. We believe the 5% elimination rate indicates that crowdsourcing can provide high quality data for such kind of tasks where participants generate language data for a given context, in both spoken and written forms.

Our first hypothesis concerns the distribution of third-person pronouns in spoken and written responses in relation to non-pronominal NPs. Based on prior findings in the literature (e.g., Fox [1987]), we expect to detect more third person pronouns in spoken than in written mode. As summarized in the previous sections, we find a statistically significant relationship between the forms of first references to the main characters and the production mode. However, frequency distribution of all NPs, which is obtained by automatic parsing, does not show a strong indication towards this direction. As a result, based on our data, we believe the first hypothesis is only partially confirmed. In order to reach conclusive evidence about all of the NPs across modes, further analysis based on accurate parsing is required.

In the analysis, we observe that participants use fewer pronouns when the story length (and hence, the anaphoric distance) is increased, in both modes. However, the gap for the average pronoun usage between 1-clause-length stories and 8-clause length stories is greater in written responses (32%) than in spoken responses (20%). We fit a generalized mixed-effects logistic regression model to these data points. The model coefficients indicate that pronoun usage is smaller in the written mode than in the spoken mode when the stimulus length is increased, as shown in Figure 5.8, which clearly illustrates that pronoun usage decreases more for the written than for the spoken mode for the long stimuli. Therefore, we infer that the anaphoric distance has a larger effect on the written mode. Exploration of pronoun usage patterns for each stimulus length reveals that pronoun usage is more common in spoken responses than in written responses for long stories, which confirms our second hypothesis and the findings of the above-mentioned comparative corpus studies.

We observe a variation in response to the stories having the same length. In our experiment, we do not control potentially explanatory variables for this variation, such as the text type of the stimulus (e.g., narrative or not), number of referring expressions in the stimulus and length of the name of the main character in the stimulus. Continuing the same line of research and conducting experiments with different control setups can can lead to a better understanding of the variation for the same length of the stimuli.

In addition to the statistical analysis, we broaden the scope of our study by performing preliminary classification experiments on the data. The task in the classification experiments is to predict the form of the first reference to the main character (i.e., pronoun or name) in a response. According to Same and van Deemter [2020], four feature categories—animacy, plurality, grammatical function, and form—that are related to the antecedent and its recency are crucial for the prediction of referential forms. Because the other potentially important features are kept constant, we believe that classification experiments on our dataset can yield valuable insights into the impact of recency. We extract 41 features from the data but achieve the best result by only employing 3 features that embed the length of the stimuli, PoS tag of the token following the first reference and the identity of the participant. Since the antecedent is always introduced at the beginning of each stimulus, the length of the stimuli and antecedent's recency are highly correlated. The other two features in the best configuration indicate that the reference's context and differences between individuals play significant roles in the prediction of referential choice. We improve over the majority baseline performance by 6.9%. We believe employing accurate syntactic knowledge can introduce additional meaningful features, to the three features with which we obtained the best result. Due to time constraints, we leave the discovery of such features based on more accurate syntactic information to future investigations.

6

Coreference Resolution of Twitter Conversations

Twitter messages represent a discourse genre that includes noisy informal language with abbreviations and purposeful typos, use of symbols such as the # and @ marks that are not seen in standard writing, unintended misspellings, etc., which makes them challenging for NLP applications. In addition to non-standard words, Twitter conversations also show peculiar phenomena of nominal coreference, such as exophoric pointers to non-linguistic content in attached visual media, and mixed pronominal references to the same entity due to the nature of multi-user conversations. Thus, tweets are a complicated genre for coreference resolution, but at the same time highly relevant for many applications that seek to extract information or opinions from users' messages. In this study, we use an existing competitive coreference resolution system [Lee et al., 2018] built with the OntoNotes corpus [Pradhan et al., 2012]. We experiment with adding coreference-annotated Twitter conversations to the training data and explore the question of whether it is sensible to blend in-domain and out-of-domain training data for the task of coreference resolution on Twitter conversations.

As we already discussed in Chapter 4, we are interested in the question of how the medium *microblog*, particularly Twitter, relates to the spoken-written spectrum with respect to coreference strategies. The quantitative corpus-based comparison described in chapter 4 indicates that signaling discourse coreference differs between spoken and written language. With this insight, we experimentally explore the question of whether coreference resolution on Twitter conversations will benefit from mode-based separation in training data, considering the spoken-like conversational style of written tweets. For this purpose, we consider the different – spoken and written – genres included in the OntoNotes corpus. We undertake experiments with in-domain Twitter data and various portions of these out-of-domain genres, and we show that carefully selecting out-of-domain genre subsets beats the straightforward "taking as much data as possible". Overall, our best configuration improves the "out of the box" performance of the system by Lee et al. [2018] on conversational Twitter data by F1 score of 21.6%.

Adaptation of coreference resolution systems to domains different from their training sets became a hot topic in the field as the systems trained on the OntoNotes benchmark exhibit poor performance in cross-domain experiments [Zhu et al., 2021, Lu and Poesio, 2021, Timmapathini et al., 2021, Bamman et al., 2020, Dakle et al., 2020, Moosavi and Strube, 2018, Ghaddar and Langlais, 2016a]. However, to our knowledge, there is no work to date specifically on the adaptation of coreference resolution to Twitter conversations.

The rest of this chapter is organized as follows: The related work in the literature is provided in section 6.1. Section 6.2 presents the data used in the experiments. Section 6.3 describes peculiar referential cases in Twitter conversations, and Section 6.4 the experiments. Section 6.5 provides various additional analyses that shed light on the domain adaptation problem followed by the conclusion and future work in Section 6.6.

6.1 Related Work

Most of the modern coreference resolution systems (e.g., Clark and Manning [2016], Lee et al. [2017, 2018], Kantor and Globerson [2019], Joshi et al. [2020]) are evaluated on the OntoNotes benchmark used in the CoNLL-2012 shared task [Pradhan et al., 2012]. Although these systems get high scores when tested on OntoNotes (e.g., Joshi et al. [2020] achieving a state-of-the-art performance of 79.6% F1 score), cross-domain experiments do not yield comparable scores. The task of adapting coreference resolution systems to different domains has gained a lot of interest in recent years due to its critical role in advanced NLP tasks such as question-answering and information retrieval.

In this section, we included recent adaptation studies from the literature targeting various written and spoken genres. We included a representative set of studies that focused on the domains from both sides of the spoken-written continuum such as edited written genres (literary texts, Wikipedia articles and biomedical texts) and unedited conversational genres (spoken dialogues). A few studies targeting social media texts and web-based communication (i.e., email conversations) are also considered in this review.

Aside from adaptation research, there exist generalizable coreference resolver implementations that aim to work effectively across a variety of domains. We included recent generalization studies which are tested on similar domains with the adaptation studies (e.g., Wikipedia articles), allowing us to compare the performance of the adaptation and generalization studies for the same domains.

Cross domain experiments for the adaptation and generalization studies are often carried out on separate datasets. However, there is empirical research that presents performance scores for different sections of the same dataset. We included these studies in this review as well since they provide substantial information for coreference resolution on the diverse domains annotated following the same scheme.

Performance scores reported for the research in this section were computed using the CoNLL scoring method [Pradhan et al., 2014], unless otherwise stated. As a result, most of the performance scores are comparable.

6.1.1 Domain Adaptation for Coreference Resolution

6.1.1.1 Literary Texts

One of the earlier studies on domain adaptation for coreference resolution is Do et al. [2015], which adapts the Berkeley system [Durrett and Klein, 2013] to narrative stories in UMIREC [Finlayson and Hervás, 2010] and N2 corpus [Finlayson et al., 2014]. Do et al. do not retrain the system but add linguistic features of narratives as soft constraints to the resolver. They perform inference via Integer Linear Programming (ILP) formulation that allows for new linguistic features not part of the original training data to be adopted as constraints by the model without retraining. The local discourse coherence, speaker-listener relations and character-naming patterns are used as soft constraints in this adaptation study. Their results indicate that character-naming constraints (e.g., An indefinite form (a woman) and a definite form (the woman) of the same noun phrase corefer.) are more beneficial than others.

Bamman et al. [2020] present a coreference-annotated corpus of English fiction works, LitBank. They first discuss the characteristics of referential relations in the literature domain. One peculiarity of the data, for instance, is the burstiness of the entities. Literature documents in the presented corpus are on average 5 times longer in terms of word length than the documents in OntoNotes. Their analysis indicates that the entities that span long text ranges exhibit bursty behavior. In other words, they tend to cluster together in tight bursts and there are sections where they are not mentioned in the text. Bamman et al. also evaluate the empirical performance of the end-to-end [Lee et al., 2017] system on this domain, by separately retraining it on three different datasets, namely OntoNotes, PreCo [Chen et al., 2018] and their literary dataset. PreCo consists of reading passages from English tests and is larger than the other datasets (12.2M tokens in PreCo vs 1.3M in OntoNotes and 210K in LitBank). The experimental results demonstrate that retraining with PreCo and LitBank produces comparable results (79.3% F1 score with LitBank and 78.8% F1 score with PreCo), which are significantly better than training with OntoNotes (72.9% F1 score). The authors argue that PreCo's comparable performance to in-domain training data, despite including out-of-domain data, demonstrates the value of large annotated datasets for cross-domain coreference resolution.

6.1.1.2 Wikipedia Articles

Ghaddar and Langlais [2016a] explore the coreference resolution adaptation problem for the Wikipedia articles in the WikiCoref [Ghaddar and Langlais, 2016b] dataset. They investigate the task of main concept resolution, which is defined as identifying in a Wikipedia article all mentions of the main concept being described. They implemented a classifier exploiting features extracted from Wikipedia markup, as well as from the external knowledge resource Freebase [Bollacker et al., 2008]. The baseline experiments are built on several rule-based and machine-learning based coreference resolution systems. Their best configuration improves the best baseline score by F1 score of 13% resulting in F1 score of 55.11%.

6.1.1.3 Biomedical Texts

Biomedical texts are one of the most studied domains for the task of coreference resolution. Lu and Poesio [2021] makes an extensive overview of the coreference resolution systems adapted to the biomedical domain. They evaluate the empirical performance of existing models on the data of BioNLP 2011 shared task [Nguyen et al., 2011]. According to the results of the evaluation, the neural model developed by Li et al. [2021] demonstrates superior performance (69.5% F1 score) compared to the previous systems. Li et al. [2021] incorporate an external knowledge base into their neural model and achieved a state-of-the-art performance for the task of coreference resolution in the biomedical domain. In comparison to the BioNLP 2011 test set, testing the same systems on another dataset from the biomedical domain, CRAFT [Cohen et al., 2017], yields lower performance. On the CRAFT dataset, for example, Li et al. [2021] still performs better than the others. However, its performance drops from 69.5% to 57%. CRAFT contains full text articles whereas BioNLP 2011 data is composed of only abstracts. As a result, the documents in the CRAFT dataset are substantially longer than the BioNLP 2011 test set, potentially making CRAFT a more challenging dataset for coreference resolution, according to the authors.

Lu and Poesio [2021] also compare the empirical performance of pre-trained language models (PLMs) for the biomedical domain. They experimented by integrating language models into the end-to-end resolver developed by Lee et al. [2018]. The tested PLMs are chosen from both biomedical domain models, such as Clinical KB-BERT [Hao et al., 2020], and Bio_ClinicalBERT [Alsentzer et al., 2019], as well as general domain models such as SpanBERT [Joshi et al., 2020]. Their experiments indicate that the setup with SpanBERT [Joshi et al., 2020] achieved the best scores on the CRAFT dataset (47.76% F1 score). Clinical KB-BERT, which integrates an external biomedical knowledge base, outperforms existing biomedical PLMs as well as the general domain language model BERT, suggesting that domain knowledge can improve biomedical domain coreference resolution systems.

6.1.1.4 Conversations

Dakle et al. [2020] investigates coreference resolution in email conversations for certain entity types (i.e., Person, Location, Organization and Digital entities). Email conversations, like the Twitter dataset we used in this work, have a thread structure where each email, aside from the one that starts the conversation, is generated as a reply to a previous email in the conversation tree. This creates challenges similar to those we discuss in Section 6.3, such as mixed pronominal references to the same entity in multi-user conversations. Dakle et al. [2020] experiment with Joshi et al. [2020]'s state-of-the-art model. When tested on the email conversation dataset, the model that was originally trained and fine-tuned on OntoNotes performs poorly (34.9% F1 score). This result is not directly comparable to the performance of the model when tested on OntoNotes as only certain entity types are annotated in the emails. However, using in-domain data to train and fine-tune the same model results in a significant gain in performance (19% F1 score).

Coreference resolution in spoken conversational domains has recently gained interest due to the CODI-CRAC 2021 Shared task [Khosla et al., 2021] which focuses on the task of coreference resolution for dialogues[1]. The benchmark used in the shared task consists of conversational sections from the ARRAU corpus [Poesio et al., 2018]. Additionally, texts from four conversational corpora (i.e., Switchboard [Godfrey et al., 1992], AMI [Carletta et al., 2006], Light [Urbanek et al., 2019] and Persuasion [Wang et al., 2019]) are annotated within the scope of the shared task using the annotation scheme of the ARRAU corpus, which is an extension of the OntoNotes scheme with additional support of various phenomena such as singletons and split-antecedents. Three aspects of anaphoric relations are included in the shared task: identity anaphora, bridging anaphora, and discourse deixis. Among the teams submitted for the identity anaphora, Xu and Choi [2021] ranks at the first place.

Xu and Choi [2021] use the end-to-end neural approach with the SpanBERT encoder [Joshi et al., 2020] as the baseline model. They integrate singleton marking into the baseline system and exploit conversational metadata (i.e., speaker encoding for each sentence in each dialogue turn). In addition, apart from the shared task benchmark, they trained their model with two external datasets: OntoNotes and BOLT [Li et al., 2016]. The external datasets are either utilized as pre-training data to initialize the system before it is further trained with the task benchmark (i.e., continued training) or blended with the benchmark data to augment the training data (i.e., joint training). Experiments show that training with external resources is effective in both configurations, even if the annotation schemes for the shared task benchmark and the external corpora differ. The setting that uses external resources for model pre-training outperforms the others and improves the baseline for all of the benchmark datasets. For example, on the Switchboard corpus, the model has a 74.5% F1 score, suggesting a F1 score of 27% improvement over the baseline.

6.1.1.5 Social media texts

Social media texts are one of the understudied linguistic genres for coreference resolution. One of the few studies on Twitter texts is that of Andy et al. [2020] which addresses the resolution of personal pronouns referring to characters/people about televised events in tweets. They develop an algorithm that takes into account the time the tweets were published as well as the context generated from other tweets on the same topic (e.g.,

[1] https://competitions.codalab.org/competitions/30312

tweets including the same hashtag). Their data is made up of tweets with the hashtag "#got" posted during the airing of an episode of "Game of Thrones" and tweets with the phrase "debate" posted during the televised presidential debate. One of the challenging cases in their data is the tweet texts containing pronominal mentions of people who are not named in the tweet. For instance, the tweet "wait where is arya did she change to **his** face" includes the word "his" as a reference to a character from "Game of Thrones", "Walder Frey", who is not mentioned in the text. For the resolution task, they first identify the characters in the events and then extract contextual embeddings of the pronoun and character tokens in tweets, using a BERT-base model [Devlin et al., 2019]. They report several results but do not describe the scoring method. Therefore, it is not clear to what extent their results are comparable to the previous scores reported for the baseline methods (e.g., Lee et al. [2017]). However, their experiments indicate that temporal information and contextual word embeddings are beneficial for the resolution of third person pronouns referring to the characters in the televised events, even for the challenging case of tweets with pronouns but no possible referent.

In another study on Twitter texts, Singh et al. [2020] investigate the problem of coreference resolution in code-mixed[2] (Hindi and English) tweets. They use the dataset provided for FIRE 2020 SocAnaRes-IL shared task[3] [Devi, 2020]. The task includes the challenges of both social media and code-mixed texts (e.g., use of emoticons, hybrid grammar, spelling variations). They implement a deep learning based approach using an encoder-decoder architecture with attention. The model uses word2vec embeddings created from the train data. Their best configuration performs with 21% F1 score. The authors advise that their study be used as a baseline for that line of research because no other work exists that addresses the problem for Hindi language.

6.1.2 Generalization in Coreference Resolution

Moosavi and Strube [2018] explore the role of linguistic features on the generalization capacity of a neural coreference resolver [Clark and Manning, 2016] and use the WikiCoref dataset as the out-of-domain test set. They integrated numerous features into the system such as the type of the mention or POS tags of the words preceding the mentions. They also mark the closest antecedents with the same head and compatible premodifiers, such as for the phrases "this new book" and "This book". However, the integration of these features does not result in a significant improvement in the resolver's performance. They extract the features automatically using the Stanford CoreNLP [Manning et al., 2014]. As a result, the feature set comprises predicted values with considerable noise, which is likely one of the causes for the lack of a significant performance gain. In order to better exploit the extracted features, they applied a pattern mining procedure to find the most informative, and hence more effective, feature values by examining all feature value combinations. The pattern mining process identified some feature values as more informative, such as specific subsets of POS tags, dependency relations, and mention categories. These values are integrated to the resolver as binary features. Incorporating these informative features improves both in-domain and out-of-domain performance. Their best result for the WikiCoref dataset (55.30% F1 score) is about three points higher than the baseline score and comparable to the model developed by Ghaddar and Langlais [2016a] (55.11% F1 score) on the WikiCoref test data despite the fact that they do not use any domain-specific features as Ghaddar and Langlais [2016a] do.

In a more recent study, Toshniwal et al. [2021] combine a collection of eight coreference-annotated datasets from different domains to evaluate the out-of-the-box performance of the coreference resolution models. Datasets differ not only in terms of text domains, but

[2] The mixing of units (e.g., words or morphemes) of one language into another language

[3] Only this reported study submitted the run for the shared task.

also in terms of annotation guidelines and document lengths covered. The three largest datasets (OntoNotes [Pradhan et al., 2012], LitBank [Bamman et al., 2020] and PreCo [Chen et al., 2018]) are used for training while the others, including WikiCoref, are used for testing and analysis. The state of the art models show poor performance when a model trained, for instance, on OntoNotes is run on unseen type of data (e.g., Wikipedia articles). They use the coreference resolver developed by Toshniwal et al. [2020] and conduct further experiments by joint training with the combination of OntoNotes, LitBank and PreCo. They suggest automated workarounds to align the differences in the annotation schemes of the training datasets. OntoNotes, for example, does not annotate singletons, while LitBank and PreCo do. They add predicted singletons to OntoNotes data to account for this disparity. Experimental results indicate that joint training improves performance on unseen data. For instance, the joint model, when tested on WikiCoref dataset, increases the F1 score by 3% over the model trained only on OntoNotes, achieving the state-of-the-art performance for WikiCoref with an F1 score of 62.5%.

In another recent study, following a similar line of research with Toshniwal et al. [2021], Xia and Van Durme [2021] explore the impact of continued training[4] (i.e., a fully trained model on a source dataset is further trained on the target dataset) on the generalization of coreference resolution systems. They use the resolver developed by Xia et al. [2020], which is a memory-efficient adaptation of the model by Joshi et al. [2020]. They utilize OntoNotes and PreCo for pre-training in different experimental setups and test the methods by further training on the subsets of several datasets including LitBank and ARRAU. They find that continued training is effective in all the setups targeting the out-of-domain texts, especially when the target data, and similarly the training data, is smaller in comparison to the pre-training data. They also claim that PreCo, which contains texts only from exam passages and is much larger in terms of data size than OntoNotes, outperforms OntoNotes as a pre-training dataset, particularly when there are only a few annotated texts for the target domain, suggesting that PreCo could be a viable alternative to OntoNotes for pre-training.

6.1.3 Cross-domain evaluation on multi-domain datasets

As seen in the papers presented in this section, cross-domain experiments are usually undertaken on distinct corpora. However, there are also studies looking at the performance differences between genres within the same corpus. For instance, regarding OntoNotes genre differences, Pradhan et al. [2013] show that a number of resolvers perform better on telephone conversations (64%) than news texts (56%) and broadcast news (59%). They evaluate these results and note which genres turn out to be the “easiest” but do not assess the possible reasons.

Beyond this result, we are aware of only one other study that examines in detail the performance differences in OntoNotes (and also in two other corpora): Uryupina and Poesio [2012] compare the performance of “domain-specific” and “generic” models, for both knowledge-poor implementations and for classifiers using hand-crafted linguistic features. They introduce a method for evaluating the similarity of the different domains in a single corpus that employs quantitative document structure indicators such as number of mentions per token and length of the longest chain. Their investigations indicate diverging results for the different domains within the same corpus (e.g., 59.5% MUC F-score for broadcast conversations as opposed to 54.5% MUC F-score for news texts in OntoNotes). They argue that choosing features carefully (i.e., the domain-specific approach) improves the model performance for larger domains, whereas they suggest combining smaller domains with other similar domains to avoid over-fitting, using the method they introduce in the paper to compute similarity.

[4] This is the term used bu the authors.

Similar to OntoNotes, OntoGUM is also a multi-domain dataset Zhu et al. [2021]. It consists of the mapped version of GUM annotations [Zeldes, 2017] onto the OntoNotes scheme, broadening the scope of OntoNotes by size and genre diversity. OntoGUM consists of texts from 12 genres including conversations and web-based communication from Reddit[5] and from YouTube Creative Commons vlogs[6]. They experiment with the rule-based *dcoref* in Stanford CoreNLP toolkit [Manning et al., 2014] and a neural coreference resolution system [Joshi et al., 2019]. Experiment results indicate that both approaches exhibit poorer performance on OntoGUM genres compared to OntoNotes. The score of the neural system degrades by 15 points on OntoGUM. The decrease in the resolution performance is more noticeable for certain genres. The neural system, for example, yields 73.3% F1 score on *vlog* data, which contains unedited texts from the web, whereas its F1 score on the *news* data is 60.6%. This result is interpreted as unexpected because *vlogs* supposedly contain texts that are substantially different from the *news* and the neural system is trained on OntoNotes, containing 30% of its data from *news* texts. The findings in this study support the necessity for testing resolution systems on a variety of corpora in order to improve coreference resolution task generalization, even for in-domain data.

6.2 Data

For our experiments,[7] we use the English portion of the OntoNotes benchmark (ONT) used as the training set in the CoNLL-2012 shared task [Pradhan et al., 2012]. It comprises texts from spoken and written genres, and contains annotations at different layers, including coreference chains. Spoken data includes telephone conversations (**tc**), broadcast conversations (**bc**), and broadcast news (**bn**); written data contains magazine (**mz**), newswire (**nw**), pivot text[8] (**pt**) and web blogs (**wb**). As shown in Table 6.1, the **ONT** corpus contains 1289K *tokens* in 2632 *documents* (in CoNLL terminology, documents are the units of independent annotation).

	docs	tokens	chains	mentions
ONT	2632	1289K	34K	152K
tc	111	81K	1931	12K
bc	284	144K	4236	18K
bn	711	172K	6138	21K
mz	410	164K	3534	13K
nw	622	387K	9404	34K
pt	320	210K	6611	42K
wb	174	131K	2993	12K
TW	185	48K	1534	6K

Table 6.1: Corpus size and basic coreference statistics (The original TW corpus was preprocessed for the alignment of its annotations with OntoNotes; therefore the annotation statistics in this table are different from the original TwiConv corpus presented in Chapter 3.)

Our second dataset is the Twitter Conversation corpus (**TW**) described in Chapter 3. The conversations are tree structures where each tweet has a parent (i.e., the tweet it is replied-to) except for the initial tweet starting the conversation. A tree can be shallow,

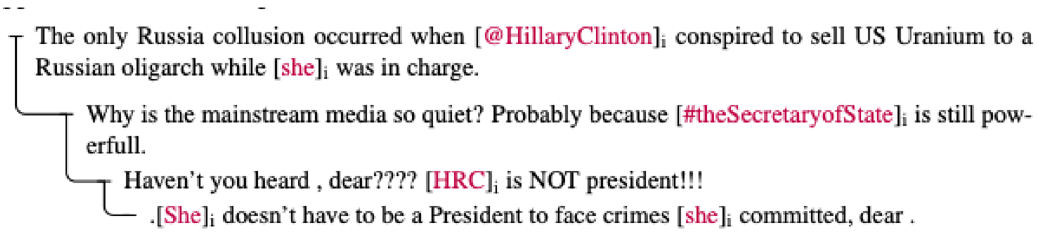
[5] https://www.reddit.com/

[6] Wikipedia definition: A form of blog for which the medium is video.

[7] Data distribution and scripts can be found at `https://github.com/berfingit/e2e-coref-twitter`

[8] Texts from the Old Testament and the New Testament

with many replies on just one level, or it can be deep when participants interact with each other across several turns. The corpus holds 1756 tweets in 185 threads, defined as a path from the root to a leaf node of a conversation tree.[9] 69% of the coreference chains in this dataset contain coreferential relations across tweets. Hence, considering the conversation context is important. We illustrate a thread structure with one example of coreference chain annotation in Figure 6.1.



The only Russia collusion occurred when [@HillaryClinton]$_i$ conspired to sell US Uranium to a Russian oligarch while [she]$_i$ was in charge.

Why is the mainstream media so quiet? Probably because [#theSecretaryofState]$_i$ is still powerfull.

Haven't you heard , dear???? [HRC]$_i$ is NOT president!!!

.[She]$_i$ doesn't have to be a President to face crimes [she]$_i$ committed, dear .

Figure 6.1: A thread sample from TwiConv

6.2.1 Corpus Homogeneity

The original TW corpus was annotated with a scheme that is marginally different from that of ONT. For a systematic comparison, we modified the TW annotations so that they conceptually align with ONT.

Only the *identity* relations are annotated in both of the corpora. Mentions building singleton chains (i.e., chains containing only 1 item) are not considered as markables in OntoNotes but TW corpus includes singleton annotations. Therefore, we excluded the singleton chains from the TW annotations, resulting in the elimination of 5301 singleton chains. After eliminating the singletons in TW, 1734 non-singleton chains with 7073 mentions remained annotated in the corpus.

In addition to the difference in handling the singletons, the annotation schemes of the two corpora are also not fully compatible in terms of their definitions of markables. For the sake of comparability of the experimental results, we aligned the type of annotated markables to the best of our ability by applying semi-automated procedures. We summarize below the main differences we observed and applied handling strategies to align with:

- In TW, predicative nouns (e.g. *This is* [*a fake account*]), and headless relative clauses having the grammatical role of a noun phrase (e.g. *A mature male kangaroo doing* [*what*] *it's built for*) are considered as markables, but this is not the case in OntoNotes. We remove the predicative noun and relative pronoun annotations in TW.

- In TW, appositions (e.g. [*His wife*], [*Florence*], *fell ill.*) are annotated separately from the preceding noun they corefer with. In the CoNLL formatted version of OntoNotes [Pradhan et al., 2013] that we use, appositions are merged with the nominals they modify (e.g., [*His wife, Florence*], *fell ill.*). Therefore, the appositive modifiers in TW are merged with the preceding coreferring noun phrase.

- Generic "you" instances are annotated in TW but not in OntoNotes. We remove generic "you" annotations from TW.

- In TW, "reflexives" are annotated as separate mentions even if they are used for focus (e.g. [*The president*] [*himself*] *said this*). However, the focus reflexives are both

[9] Only the longest path has been used from each tree, so there is no redundancy in the data.

annotated as separate markables and also part of the span of the preceding coreferring noun phrase in OntoNotes (e.g. [*The president* [*himself*]] *said this*). Therefore, the focus reflexives in TW are added to the span of the preceding coreferring noun phrase.

- In OntoNotes, verb mentions are annotated if they corefer with a nominal mention (see Example 6.1). However, no verb mention is annotated in TW. We remove the verb mentions in OntoNotes for one experiment setting (see Section 6.5 for details) but not for the main experiment for the sake of comparability with the reported scores in previous studies.

  (6.1) Sales of passenger cars [grew]$_i$ 22%. [The strong growth]$_i$ followed year-to-year increases.

If the removal of a mention makes the remaining chain a singleton (i.e., only 1 mention left in the chain), the whole chain is removed from the annotations, as no singleton chains are allowed in the OntoNotes scheme. As a result of the alignment process, 12% of the chains (200 out of 1734) and 10% of the mentions (719 out of 7073) were eliminated from the TW corpus.

6.2.2 Pre-processing

Apart from the alignment of the annotations, we only perform two trivial preprocessing steps on the TW dataset:

- We normalize the parentheses, namely left and right bracket tokens, into '-LRB-' and '-RRB-', respectively.
- We converted all smiley and emoji tokens into the strings of "%smiley" and "%emoji", respectively.

6.3 Coreference Phenomena in Twitter Conversations

As we discuss in detail in chapter 4, Twitter texts contain non-standard referential expressions. As seen in the examples below, these non-standard mentions include usernames, hashtags (either integrated or not integrated into the syntax of the tweet), emojis, and links.

(6.2) .. [@SomeUser] just said twice that.. *("username" as a mention)*

(6.3) this doesn't pass [the #smelltest] *("hashtag" integrated to the syntax and used as a part of a mention)*

(6.4) .. [#IranianProtests] THE DEMOCRATS AND LINDA SARSOUR HATE THESE PROTESTS *("hashtag" not integrated to the syntax and constitutes a separate mention)*

(6.5) [🐍] are fools ... *("emoji" as a mention)*

(6.6) If crashing, please refer to this: **[https://exampleurl.com]** *("link" as a mention)*

As illustrated in example 6.5, graphical elements, such as emojis, can also be utilized as referential expressions in tweets. However, we are only interested in coreference resolution on textual components. Therefore, we substitute all emojis with a text string, as explained in Section 6.2.2. Despite the fact that it eliminates the distinctions between the emojis, this step was essential for the coreference resolver we use to process our data to work effectively.

In addition to the above-mentioned non-standard nominal mentions, Twitter conversations exhibit peculiar phenomena related to nominal coreference, either as a consequence of the user interface, character limitation in the texts[10] or popular conventions applied by the users. The following illustrates the situations that we have run into during the annotation process.

1. **Non-aligning replies** A potential complication in any approach to analyzing Twitter conversations from a discourse perspective is possible mismatches between the replied tweet and the actual relation based on the contents of the tweets: In certain Twitter UIs, it may well happen that a user reads a sequence of related tweets, hits "reply" to tweet X, but then in fact responds to a different tweet Y in the neighborhood of X. We observe a few clear cases in our threads. In general, these cases can be hard to detect automatically, and it is not possible to reliably estimate the frequency of the problem solely on the basis of our relatively small sample. Hence we leave a deeper investigation for future work.

2. **Multi-user conversations** In addition to using proper names, speakers can refer to one another using pronouns. In multilogue, it is possible that third-person pronouns *he/she* refer to conversation participants which can result in chains with first, second, and third person pronouns referring to the same entity. The tweets shown in 6.7-6.9 construct a thread segment with three users @realDonaldTrump, @user1, @user2[11]. In the conversation, the pronouns "I", "you" and "he" are used to refer to the writer of the first post (@realDonaldTrump) as indicated by the subscripts on relevant pronouns.

   (6.7) 1: **@realDonaldTrump:** [I]$_{\mathrm{i}}$ 've had to put up with the Fake News from the first day [I]$_{\mathrm{i}}$ announced that [I]$_{\mathrm{i}}$ would be running for President. Now [I]$_{\mathrm{i}}$ have to put up with a Fake Book, written by a totally discredited author. Ronald Reagan had the same problem and handled it well. So will [I]]$_{\mathrm{i}}$!

   (6.8) 2: **@user1:** @realDonaldTrump Stay strong. [You]$_{\mathrm{i}}$ are our hero. I'm so proud to call [you]$_{\mathrm{i}}$ MY president. As an educated female, I would be the first to stand up for [you]$_{\mathrm{i}}$. I'm so tired of the fake news.. [..]

   (6.9) 3: **@user2:** @user1 @realDonaldTrump [He]$_{\mathrm{i}}$ can quote things out [his]$_{\mathrm{i}}$ mouth and you hear [him]$_{\mathrm{i}}$. Come back two days later and say, fake news. [His]$_{\mathrm{i}}$ base will agree with [him]$_{\mathrm{i}}$.[..]

3. **Exophoric reference** This concerns the use of first and second person pronouns as also mentioned as a natural result of multi-user conversations above as illustrated in the conversation between @user1, @user2 and @user3 in 6.10-6.12. Resolving such coreference chains requires knowledge of tweet authors and of the *reply-to* structure.

   (6.10) 1:**@user1**: [[my]$_a$ aunt]$_i$ won't eat anything.

   (6.11) 2:**@user2**: @user1 [[my]$_b$ aunt]$_j$ eats everything.

   (6.12) 3:**@user3**: @user1 @user2 hope [[your$_{a/b?}$ Auntie]$_{i/j?}$ picks up soon.

[10] At the time the data collected, the character limit was 280 characters for the tweets.

[11] The usernames of @user1 and @user2 are anonymized for privacy.

Antecedent in the attached media:	56
Antecedent in the quoted tweet:	28
Antecedent in the attached link:	30
Antecedent inferred by world knowledge:	120

Table 6.2: Exophoric reference statistics

Furthermore, Twitter allows users to insert images, videos and URLs into their tweets. As a result, tweets are multi-modal. It is also possible to quote (embed) a previous tweet and comment on it. For anaphora, this means that antecedents can be entities in embedded images, videos, and even in the text of referred URL or an embedded tweet, or its author. We annotate these anaphors where the antecedent is out of the current linguistic domain (i.e., the text of the tweet or its preceding tweets) as exophora, using the categories given in Table 6.2. As the numbers in the table show, in most cases of exophora, the antecedents can be found in the attached pictures. For example, in the following conversation, “her” in the second tweet refers to an entity in the embedded picture in the first tweet.

(6.13) 1:**@user1**:Few more of me on the way to work had to get the Train into day as Toms car in the Garage so he had to take mine did I sit opposite you today on the train if I did did u notice my stocking Xxx PICTURE_URL

(6.14) ..

(6.15) 4:**@user2**:@user1 i know i would have enjoyed the view! make eye contact, gesture **her** to show me more

A final category of exophoric reference results from Twitter’s listing the top keywords or hashtags being currently discussed (“trending topics”) in the UI. For example, 6.16 is a tweet that appeared after the 2017 Golden Globe awards:

(6.16) Come onn! How can she be a president?!

Most probably, *she* in 6.16 refers to Oprah Winfrey, as her possible presidential candidacy was a trending topic emerging from the ceremony. In such cases, we annotate *she* as an exophoric type of pronoun and assign the attribute “antecedent can be inferred by world knowledge”.

4. **General Twitter challenges** Finally, we mention some of the phenomena that are well-known problems in Twitter language, focusing here on those that can have ramifications for coreference resolution.

   - Typos affecting referring expressions:
     She not qualified to ***he*** *president why?*
   - Name abbreviations are frequent. E.g., *Barack Obama* can be referred to as *BO, O.*, etc.
   - Missing apostrophe in contracted copula:
     Hes *my best.*
   - Intentional misspellings:
     Its ***himmm*** *who does it.*

- Deviations in the usage of upper and lower case.
  See Example 6.4 where the whole tweet content is written in upper case.
- Frequent elision, e.g., of subjects.
- Missing punctuations.

6.4 Experiments

We aim to improve automatic coreference resolution on conversational Twitter texts. We do not address the challenges of this genre of data specifically, but instead retrain an existing neural coreference resolver with in-domain data and evaluate the outcome. However, in order to ensure that both the train and test sets contain a balanced amount of the phenomena described in Section 6.3, we consider the distribution of automatically recognizable Twitter challenges (see Section 6.4.1 for the details) while dividing the corpus into its subsets. Similar to some of the research (e.g., Xu and Choi [2021]) reported in section 6.1, we experiment by jointly training the resolver with blending out-of-domain data and in-domain Twitter data.

While designing the experimental setups, we were inspired by the suggestion of Uryupina and Poesio [2012] about combining "similar" domains in the training data for small datasets. In Chapter 4, we look at the similarities between genres and the language modes in terms of quantitative features of referential expressions. In our analysis, we demonstrate that spontaneous spoken conversations (e.g., telephone conversations) and edited written genres (e.g., news) exhibit clearly different patterns. The position of the Twitter conversations in spoken-written continuum differs with respect to the examined feature. For instance, in terms of the distribution of personal pronouns, Twitter conversations resemble spoken conversations, but in terms of average anaphoric distance, computed in terms of noun phrases between anaphors and their antecedents, they are closer to edited written texts. Therefore, we empirically investigate whether coreference resolution on Twitter conversations will benefit from mode-based separation in external training data. To explore this research question, we combine the in-domain Twitter data with out-of-domain data from different sections of OntoNotes and ran a series of experiments with an end-to-end neural coreference resolver to find the best combination.

For our experiments, we use 'e2e-coref' [Lee et al., 2018], an updated version of the end-to-end neural coreference resolver of [Lee et al., 2017]. It introduced a refined approach based on differentiable approximation to higher-order inference and ELMo embeddings [Peters et al., 2018] for span scoring, which significantly improves the performance on English ONT. The approach achieved 73.0% F1 score, representing the 2018 state-of-the-art. It constitutes a basis for several more recent state-of-the-art models, including SpanBERT Joshi et al. [2020] which achieves better performance than e2e-coref on the OntoNotes dataset (79.6% F1 score). While there are better state-of-the-art results, our study focuses on the model by Lee et al. [2018] due to computational resource constraints[12].

Since e2e-coref expects its input data in CoNLL format, we convert the Twitter conversation dataset to CoNLL format. We also integrate author information for tweets in the CoNLL data, similar to existing speaker information for spoken data in OntoNotes.

[12] The decision to use Lee et al. [2018] in our work is mainly motivated by its competitive to state-of-the-art performance with no need of extensive computational power to be retrained, which allows for experimenting even without extensive resources in different setups and on longer documents. Our experimental setups are run on a GPU GeForce GTX 1080 8Gb. In comparison, the aforementioned system by Joshi et al. [2020] requires at least 32Gb of graphic card memory. According to Joshi et al. [2019], Bert-large models improve over other models especially in pronoun resolution and lexical matching, which are also relevant for our task on the Twitter corpus. Therefore, we would expect that Joshi et al. [2020]'s approach to perform better in our setup as well. However, we see this question as peripheral to the present study and thus regard it as an interesting task to explore in the future.

6.4.1 Test set

Our main goal is to examine how different training set configurations affect the coreference resolution performance on Twitter data. In order to achieve informative results with this non-homogeneous and highly variable dataset, we select a representative test set not via random sampling, but through statistical analysis of three features: **number of tokens, chains** and **mentions** per document. To meaningfully represent threads of all lengths, we choose the documents where these variables are situated either on the median, or in the first and fourth quartiles of the respective distribution, while omitting obvious outliers (see Figure 6.2). Because of the linear correlation of the parameters shown on Figure 6.3, we only select the documents where all three are in a similar range of their distributions.

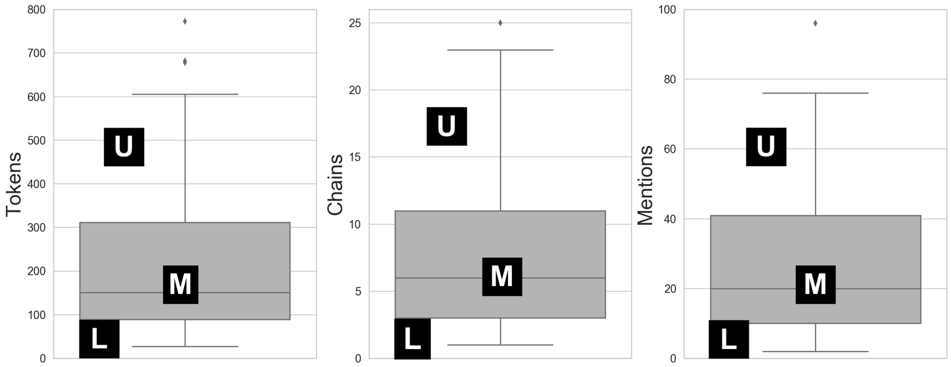


Figure 6.2: Distribution of the three considered parameters. U, L, M marks the forth (upper), first (lower) quartiles, and median respectively.

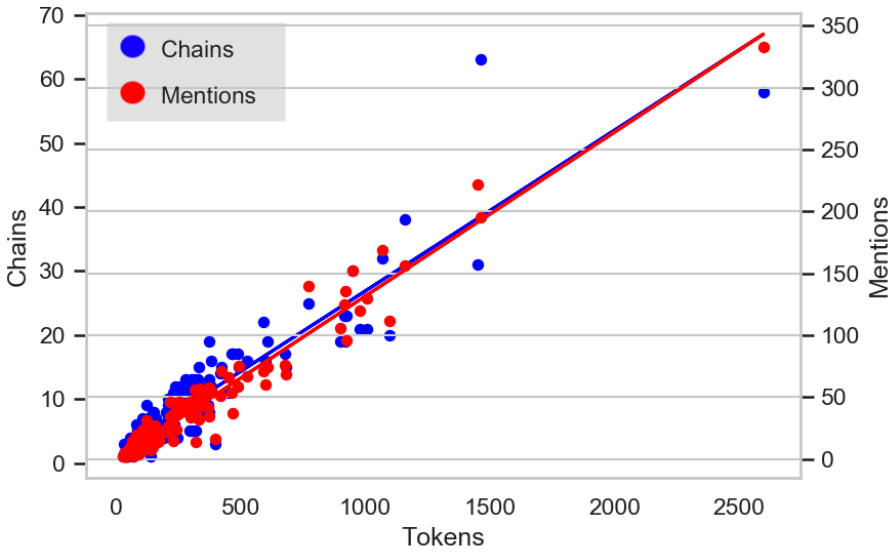


Figure 6.3: Each blue data point represents the chain vs token count for each document, while red points denote mention vs token information of the same documents.

Among the pre-screened files, we manually inspect each document, marking features of the annotated mentions (person, number, gender) as well as Twitter phenomena (hashtags, user names, pronouns with typos etc.). With this information, we exclude threads that do not contain sufficient coverage and variability of the phenomena in focus. The final distribution is shown in Table 6.3.

	Tokens	Chains	Mentions
train	44885	1411	5946
test	3260	123	408

Table 6.3: Twitter train/test distribution

6.4.2 Baseline Experiments

For evaluation, we utilize the official CoNLL-2012 scoring scripts[13] [Pradhan et al., 2014], measuring the precision, recall and F1 scores for MUC, B-Cubed and CEAFe metrics as well as the unweighted average of these scores, which is often referred to as CoNLL score.

After we successfully reproduce the published e2e-coref results, we measure how a model trained on ONT performs on our Twitter test set (**Experiment A**) (See the experimental setup indicating the quantitative features of training sets for all the experiments in Table 6.4). The resulting 45.18% F1 score (see Table 6.5) is almost F1 score of 28% lower than the result reported on the official ONT test set (actual number 73% F1 score).

A second baseline results from using only the TW twitter corpus as training data, which leads to 60.8% F1 score(**Experiment B**). Although this model is based on a rather small training set, it significantly improves on Experiment A and highlights to the difference between in-domain and out-of-domain training.

Experiment	Tokens	Chains	Mentions
A - ONT	1289K	34K	152K
B - TW only	44.8K	1.4K	5.9K
C - TW+ONT	1333.8K	35.4K	157.9K
D - TW+spoken genres	269.8K	7.5K	35.9K
E - TW+written genres	269K	5.8K	22.8K

Table 6.4: Experimental setups with respect to training data combinations

6.4.3 Main Experiment

We investigate the effects of selecting training (sub-)sets in the resolution of the coreferential entities in TW corpus. Noting that the presence of Twitter data in the training set is beneficial, for **Experiment C** we merged ONT and TW, with the latter constituting 3.35% of the total size (see Table 6.4). The results show not only a performance increase of 17% in comparison to Experiment A, but also a 2% gain over Experiment B, demonstrating that combining both ONT and TW can lead to better learning. To study this in more detail, we measure how the performance on the test set reacts to training on different subsets of ONT. We coarsely separate **spoken**, spontaneous language from **written** or edited texts.

In **Experiment D**, the training set consists of Twitter and ONT's conversational spoken genres, viz. broadcasts conversations and telephone conversations. In this dataset, the proportion of Twitter data in the training set is 16.6% as opposed to 3.35% in the original set. We observe an increase in overall performance of 4.3%, indicating that the

[13] https://github.com/conll/reference-coreference-scorers

Experiment	**Rec.**	**Prec.**	**F1**	**Rec.**[1]	**Prec.**[1]	**F1**[1]	**Rec.**[2]	**Prec.**[2]	**F1**[2]
MUC									
A - ONT	38.24	55.89	45.41	35.74	51.36	42.15	41.05	66.47	50.75
B - TW only	56.84	74.65	64.54	50.95	70.89	59.29	-	-	-
C - TW+ONT	60.35	71.07	65.27	46.38	67.77	55.07	62.8	73.06	67.54
D - TW+spok	62.1	77.97	68.41	47.9	75.44	58.6	61.75	72.72	66.79
E - TW+writ	60.35	71.36	65.39	54.75	69.23	61.14	62.45	73.85	67.68
B^3									
A - ONT	35.14	56.02	43.18	33.19	51.68	40.42	37.21	66.78	47.79
B - TW only	51.64	68.77	58.99	46.31	63.52	53.57	-	-	-
C - TW+ONT	55.95	66.02	60.57	44.58	63.04	52.23	58.29	68.97	63.18
D - TW+spok	58.25	74.16	65.25	46.46	71.45	56.31	57.16	68.48	62.31
E - TW+writ	55.19	63.9	59.23	49.28	60.4	54.28	59.24	68.85	63.68
CEAFE									
A - ONT	44.5	49.76	46.98	43.26	47.59	45.32	49.13	61.04	54.44
B - TW only	50.97	69.66	58.87	44.54	65.96	52.96	-	-	-
C - TW+ONT	56.68	67.68	61.69	50.0	65.48	56.71	59.29	70.12	64.25
D - TW+spok	61.81	71.06	66.12	53.94	68.2	60.24	59.64	64.92	62.17
E - TW+writ	52.4	67.85	59.13	46.01	64.06	53.55	58.14	67.47	62.46
Average									
A - ONT	39.29	53.89	45.18	37.39	50.21	42.6	42.46	64.76	50.99
B - TW only	53.15	71.025	60.8	47.27	66.58	55.27	-	-	-
C - TW+ONT	57.76	68.25	62.51	46.9	65.43	54.67	60.12	70.71	**65.0**
D - TW+spok	**60.72**	**74.39**	**66.8**	49.43	**71.69**	**58.3**	59.51	68.7	63.76
E - TW+writ	55.98	67.7	61.25	**50.01**	64.56	56.32	59.94	70.05	64.60

Table 6.5: The experimental results are shown in terms of MUC, B-cube and CEAFE metrics. The average displays the mean of these scores which we use to evaluate the performance of the experiments with respect to each other. (F1[1] , F1[2] are calculated after removing first and second person pronouns, and verb mentions respectively. They are discussed in Section 6.5.)

written genres may rather be detrimental rather than beneficial. However, it is not entirely clear whether the improvement results from excluding the written genres or from increasing the proportion of Twitter data.

To answer this question, we proceede to **Experiment E**, which combines the proportion of Twitter data present in Experiment D with documents from written genres. We choose newswires (nw) and magazines (mz). Experiment E scores F1 score of 61.25%, which is 5.5% F1 score lower than Experiment D. This result may partly be due to the sparsity of written data, which contains a smaller amount of chains and mentions in the written genre documents (cf. Table 6.4). Nevertheless, it still indicates an advantage of the spoken portion of ONT over the written one.

6.5 Additional Analysis

To gain further insight into the adaptation of coreference resolution to Twitter conversations, we quantitatively and qualitatively compare the results of the best-performing setup (D) to the baselines (see Table 6.6).

Mention length For all experiments, the average token length of mentions additionally predicted by the system (spurious predictions) is significantly longer ($p \leq 0.05$) than those of the correct predictions. The higher the proportion of ONT training data (whose mentions are on avg. 0.72 tokens longer than in TW), the longer those predictions are. Hence there is a tendency to select longer spans (especially when training on ONT) by the model, but these produce more errors.

Twitter-specific tokens Hashtags and usernames caused many errors in Experiment A. In reply tweets, usernames are inserted at the beginning, so the majority of the usernames at the beginning of the tweets are not part of the syntax and have not been annotated. Table 6.6 shows that many of those names are incorrectly detected as mentions, while hashtags are completely ignored. With Twitter training data being included in Experiment B, identification of Twitter-specific tokens is more successful. Tweet-initial usernames are ignored as mentions and a number of username and hashtags are now correctly predicted. Experiment D shows further improvements for syntactically-integrated hashtags, but usernames or non-integrated hashtags still remain unresolved.

Pronouns Although they are relatively evenly distributed in the gold annotations, more third person pronouns are resolved than first and second person pronouns in Experiment A, resulting in an overall F1 score of 0.769%. In Experiment B with Twitter training data, which is rich in pronouns, pronoun performance improves for first, and to a greater extent second person pronouns, and remains the same for third person pronouns, improving the F1 score to 0.917%. In Experiment D, pronoun performance is marginally worse (0.905%).

As the entire training data in B and D is conversational, which contains many first and second person pronouns, we repeat all experiments after removing those chains containing only first and second person pronouns. This is done to make sure that the improvement is not exclusively caused by the easy detection of deictic pronouns. The results are in column $F1^1$ in Table 6.5. While deictic pronouns have a major impact on F1, we still see improvements over the baseline for all experiments but C, meaning that generally, the detection of other anaphoric expressions improves as well.

Verb annotations Verb mentions exist in ONT if they corefer with a nominal mention Pradhan et al. [2007], but they are not annotated in TW. Thus, four predicted verb mentions in Experiment A, of which two are correctly linked with the demonstrative pronoun *that*, are counted as erroneous predictions. After adding training data from TW in Experiment D, no verbal mentions are predicted. To further analyze the influence of this annotation difference, we also run all experiments after removing the verbal annotations from ONT. This reduces mentions by 2.4% and chains by 3.6% in the whole OntoNotes benchmark ONT. The written and spoken data portions included in experiments D and E are affected by these changes at different rates. In spoken data, mentions are reduced by 3.1% and chains by 5%, while in the written part 1.7% of the mentions and 2.4% of the chains are eliminated. Column $F1^2$ in Table 6.5 shows the experimental results. While training with only spoken genres outperformed more written dominant training data in previous experiments, Experiment D proves otherwise by resulting in the worst results. One reason why the exclusion of verbs could have a negative effect on Experiment D might be the greater loss of training annotations in spoken data compared to written data. These variations motivate us to look further into the specific effects of different training data combinations and how verb annotations (both generally and depending on text genres) affect an otherwise purely nominal coreference resolution task.

Chain Linking The last section of Table 6.6 shows that Experiment B improves the number of correctly predicted chains compared to Experiment A, and it further increases in Experiment D, almost doubling from Experiment A. The number of partially correctly predicted chains also increases over the experiments A-E, and the number of missed entities (cases where not a single mention of an entity is predicted) is reduced by 51.3%. Notably, chains consisting only of identical strings profit the most from the combined training set in D.

	Gold	**A**	**B**	**D**
Predicted Mentions	408	305	307	334
Usernames	8	51	6	5
tweet-initial	1	44	0	0
Hashtags	11	0	4	5
Correctly Predicted	408	218	265	293
Avg. #tokens	1.64	1.41	1.13	1.18
Pronouns	219	149	199	194
1st person	57	38	53	50
2nd person	64	26	63	62
3rd person	68	60	61	59
Usernames	8	6	5	5
tweet-initial	1	1	0	0
Hashtags	11	0	3	5
Predicted Chains	123	110	90	107
Correct Chains	-	18	27	37
Partially Correct[14]	-	10	11	14
Missed Entities	-	39	32	20

Table 6.6: Properties of original gold annotations and predicted mentions and chains in the experiments A, B and D

6.6 Conclusion and Future Directions

We show that the performance of a state-of-the-art "standard" coreference resolution system run on Twitter conversations can improve by 21.6% F1 score by joint training with out-of-domain and in-domain data. In fact, even small amounts of added in-domain data in the training can have an impact. Similar to the relevant research [Bamman et al., 2020, Xu and Choi, 2021], our experiments confirm that including out-of-domain data in the training set is beneficial. Further, for the out-of-domain training data (ONT), the choice of genres, and relatively the mode, can make a bigger difference than simply scaling the size of the dataset.

Our additional analyses consider two more variants of the main experiment design: While all results given in Table 6.5 indicate that adding Twitter data to the training set improves the performance significantly, the best combination of in-domain and out-of-domain data can depend on specific factors discussed in section 6.5. Also, we show that improvements stemming from including Twitter training data do not result solely from the large proportion of 1st and 2nd person pronouns (as one might hypothesize). We additionally report that the ratio of correctly recognized Twitter-specific tokens, such as hashtags and usernames, as referential expressions is considerably improved by the addition of in-domain training data, indicating some of the challenges presented in section 6.3 may be resolved even without directly addressing them in a coreference resolver implementation. However, existing non-standard linguistic usage forms in Twitter texts, such as intentional misspellings (e.g., "himmm" instead of "him") and missing apostrophes in

[14] All gold mentions found, but also spurious mentions.

contracted forms point to the need for applying more advanced automated pre-processing procedures for improving the coreference resolution of Twitter conversations. Because we specifically examine the effects of different training data combinations on coreference resolution, rather than assessing performance changes as a result of alternative pre-processing steps, we leave the application of advanced pre-processing procedures for future study.

Finally, we test the effect of removing verb mentions from ONT, which exhibits different patterns than other setups regarding the best combination of training data. The experimental outcome encourages deeper exploration of training data arrangements in terms of these features.

In future work, we propose to focus more on the specific kinds of training data portions and examine the effect of spoken versus written modes, and that of formal versus informal language (which need not necessarily coincide). Furthermore, as Xu and Choi [2021] suggest that continued training can be more effective than joint training, it can be considered, in a future work, as an additional experimental setup with the same data portions we used for joint training in our experiments.

7

Extending the TwiConv Corpus: Coherence Relations

When interpreting a discourse, comprehenders do not evaluate each clause or sentence separately. Rather, they establish semantic connections between them. These connections are referred to as coherence relations [Hobbs, 1979, Sanders et al., 1992]. To illustrate the concept of coherence, consider the following minimal pair examples adapted by [Stede, 2011, p.79] from Hobbs [1979]:

(7.1) John took a train from Paris to Istanbul. He has family there.

(7.2) John took a train from Paris to Istanbul. He likes spinach.

7.1 has a natural flow that is straightforward to grasp. However, perception of 7.2 as a coherent text is less straightforward and requires extra explanation to be made. For instance, for the same example Hobbs [1979] suggests the scenario that perhaps the spinach crop failed in France and Turkey is the closest country in which spinach is available. Under this assumption, one can now infer a causal relation between the sentences analogous to the one in passage 7.1 and as a result the passage 7.2 is more natural. If readers are not able to construct such relations between the clauses and sentences, they will be unable to fully comprehend the text. In the passages above, the establishment of coherence which is crucial for natural language understanding is achieved via coherence relations between sentences.

According to Hobbs [1979] pronoun resolution will follow as a by product of inferring coherence relations. Hobbs [1979] argue that coherence relations can override the other factors and heuristics influencing the pronoun resolution, such as the subject-preference heuristic (i.e., among other things the subject is favored over the object in pronoun resolution) [Hobbs, 1976]. [Chiriacescu, 2011, p.33] uses following examples (adapted from Kehler et al. [2008]) to illustrate the validity of this hypothesis:

(7.3) Bush narrowly defeated Kerry, and as a result he took some days off. [he=Kerry]

(7.4) Bush narrowly defeated Kerry, and then he took some days off. [he=Bush]

When the clauses are connected by a causal relation (as in 7.3), the pronoun in the second sentence refers to the direct object (Kerry) of the first sentence rather than the subject, whereas when the clauses are connected by a temporal relation (as in 7.4), the subject pronoun “he” in the second clause more likely refers to the subject (Bush). Following in the tradition started by Hobbs [1979], works like Kehler [2002], Kehler et al. [2008] examined how coherence relations and coreference interact. Kehler et al. [2008] conduct controlled experiments to see how coherence relations affect pronoun resolution when compared to the previously indicated semantic and syntactic restrictions of grammatical role parallelism, thematic roles, implicit causality, and aforementioned subject preference. They argue that "the coherence-driven analysis can explain the underlying source of the biases and predict in what contexts evidence for each will surface" [Kehler et al., 2008, p.1], offering experimental evidence for the impact of coherence relations on pronoun resolution. These findings suggest that researching coreference and coherence interactions together is essential.

There are currently a number of frameworks in use for labeling coherence relations in discourse such as Rhetorical Structure Theory (RST, Mann and Thompson [1988]), the Cognitive Approach to Coherence Relations (CCR, [Sanders et al., 1992]), Segmented Discourse Representation Theory (SDRT, [Asher and Lascarides, 2003]), and the Penn Discourse TreeBank (PDTB, [Prasad et al., 2008]). Some of these approaches make strong assumptions about the structure of the discourse. RST, for example, represents discourse as a tree and SDRT as a graph. PDTB, different from those, makes no assumptions about discourse structure and is more concerned with capturing local coherence between adjacent or textually proximate units than capturing the overall structure of the discourse. As a result, PDTB relations are also known as shallow discourse relations, and shallow discourse parsing is the process of parsing the discourse in terms of PDTB relations. Because PDTB makes no assumptions about the general discourse structure, it is regarded as a theory-neutral approach to coherence relation analysis. On top of our coreference-annotated corpus of TwiConv, we added PDTB-style annotations. The structural peculiarities of social media conversations have not yet been thoroughly investigated, therefore we preferred to work with the flexible annotation style of PDTB because it does not make assumptions about the general structure of the texts (i.e., whether it is a tree or a graph). As a result, considering only the local relations accelerates the annotation process since the annotators do not need to consider the global structure for each annotation unit while making judgments. In addition, the PDTB annotation tool [Lee et al., 2016] provides a stand-off representation of annotations (i.e., Annotations stored separately from the text, represented with character indices) that can be merged with other stand-off annotation layers such as the coreference annotations we have.

7.1 Background: Penn Discourse Treebank

PDTB refers to both the largest corpus, composed of news texts, annotated for discourse relations[1] and the framework describing the annotation of coherence relations. The main purpose of PDTB-style annotation is to identify two (mostly consecutive) arguments Arg1 and Arg2 which are semantically related. This relation can be constructed via explicitly expressed discourse connectives (i.e., an explicit relation) or can be inferred implicitly (i.e., an implicit relation). In case of the presence of a linking discourse connective, 'Arg2' denotes the argument that is syntactically integrated with the connective in the same sentence, whereas 'Arg1' refers to the "external" argument. Arg1 and Arg2 are usually adjacent (see example[23] 7.5). However in some circumstances, Arg1 may be non-adjacent (see example 7.6). PDTB distinguishes more relation types beyond explicit and implicit relations, such as EntRel, AltLex, and Hypophora (for the details see Webber et al. [2018]). In our annotations, we annotate explicit, implicit and hypophora relations, with the latter referring to the text's question-answer pairs. Two instances of hypophora relations are illustrated in 7.7 and 7.8, which represent the situations where the two arguments (question

[1] We use the terms "coherence relations" and "discourse relations" interchangeably.

[2] Examples are taken from PDTB dataset by [Stede, 2011, p.101]

[3] In the examples given in this section, Arg1 is marked by *italic letters*, Arg2 by **bold letters** and connectives by underlining.

and response) are from distinct Twitter users and from the same user, respectively.

(7.5) *Drug makers shouldn't be able to duck liability* because **people couldn't identify precisely which identical drug was used.**

(7.6) *France's second-largest government-owned insurance company, Assurances Generales de France, has been building its own Navigation Mixte stake* currently thought to be between 8% and 10%. Analysts said **they don't think it is contemplating a takeover**, however, and its officials couldn't be reached.

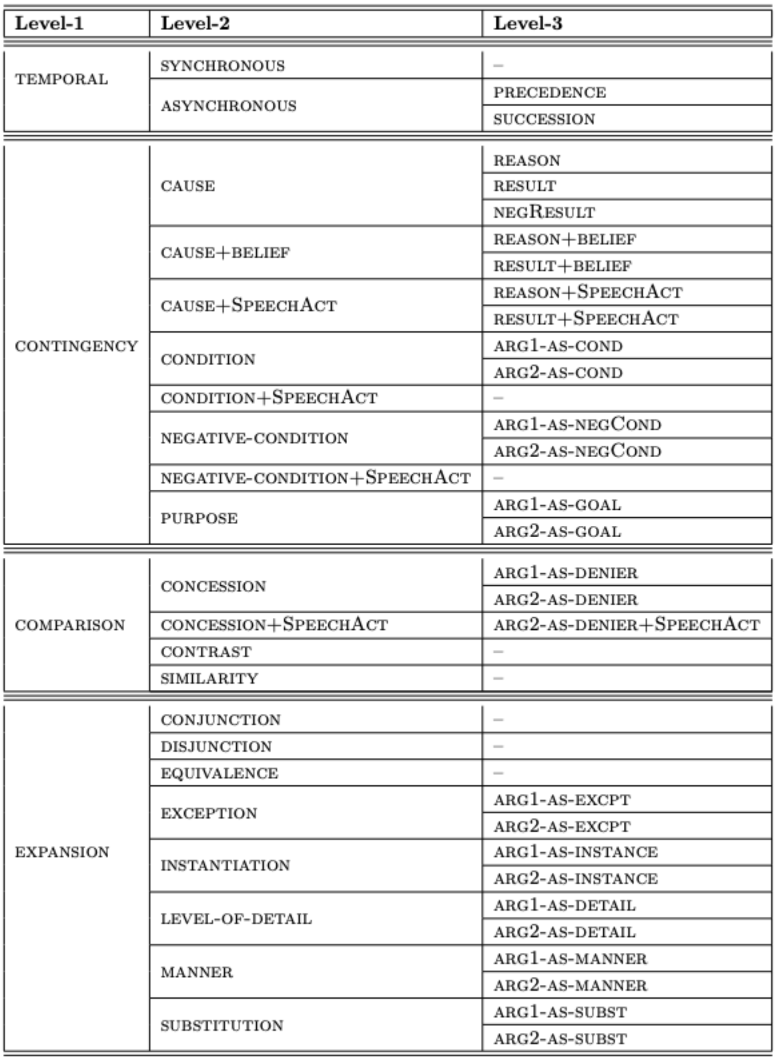
(7.7) User 1: *why you angryyy?*
User 2: **I just don't like teams stacking up like that**

(7.8) User 1: *Why don't you have outrage for this?* Oh I forgot. (Implicit=Because) **It's another failed attempt to deflect.**

We followed the PDTB-3 scheme in our annotations. PDTB-3 framework uses a 3-level hierarchy for the relation sense labels, where at the top level is the "class" label distinguishing "between EXPANSION (one clause is elaborating information in the other), COMPARISON (information in the two clauses is compared or contrasted), CONTINGENCY (one clause expresses the cause of the other), and TEMPORAL (information in the two clauses is temporally related)" [Stede, 2011, p.101]. Level-2 and level-3 in the label hierarchy represent the fine-grained labels refining the semantics of the class labels. The complete sense hierarchy for the PDTB-3 is shown in Figure 7.1.

As previously stated, PDTB dataset is composed of written texts of news (from Wall Street Journal). There exist studies applying the PDTB framework to spoken text types such as telephone conversations and broadcast interviews [Rehbein et al., 2016], a variety of formal and informal conversations, interviews or political speeches [Crible and Cuenca, 2017], and help desk dialogs [Tonelli et al., 2010, Riccardi et al., 2016]. These studies show that the use of discourse connectives and discourse relations differs significantly between written and spoken data. There is evidence that the types of relations and connectives found in Twitter conversations differs from written news text and reflects some features of spoken conversations Scheffler [2014], Scheffler and Stede [2016]. For instance, Scheffler [2014] shows that the German causal connectives are much less frequent in tweets and in spoken German than in the written texts, indicating tweets show a more oral-style characteristic than the written-style in this sense. Following this line of research, our new dataset would be useful for exploring the genre-based differences in shallow discourse between monologues and conversations, in addition to serving as a resource for investigating the interplay of coherence and coreference relations.

PDTB-3 is the extended version of PDTB-2 dataset where the majority of the newly annotated relationships occur intra-sententially [Prasad et al., 2018]. We annotate the intra-sentential explicit relations (see the Appendix D) in addition to the inter-sentential explicit relations. However, because annotation of implicit relations is a labour-intensive effort, we first focused on the inter-sentential implicit relations and did not include the annotation of intra-sentential implicit relations in this version of our corpus. As a result, the distribution of relation frequencies (number of explicit vs implicit relations) between our data and the PDTB-3 corpus may not be directly comparable. However, quantitative features concerning the relative distributions for each relation type, such as percentage of EXPANSION and CONTINGENCY types for the implicit and explicit relations in PDTB and TwiConv, can be deemed comparable and can provide insight into the distinctions between the genres of news and Twitter conversations.



Level-1	Level-2	Level-3
TEMPORAL	SYNCHRONOUS	–
	ASYNCHRONOUS	PRECEDENCE
		SUCCESSION
CONTINGENCY	CAUSE	REASON
		RESULT
		NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF
		RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT
		RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND
		ARG2-AS-COND
	CONDITION+SPEECHACT	–
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND
		ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	–
	PURPOSE	ARG1-AS-GOAL
		ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER
		ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
	SIMILARITY	–
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT
		ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE
		ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL
		ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER
		ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST
		ARG2-AS-SUBST

Figure 7.1: PDTB-3 sense hierachy (taken from [Webber et al., 2018, p.17]

7.2 Data

7.2.1 Annotation of Coherence Relations

The TwiConv corpus has been annotated for PDTB-style coherence relations. We applied the specifications in PDTB 3.0 annotation manual [Webber et al., 2018], if possible, and introduced our own specifications if data requires further customization of the PDTB annotation rules as presented in our guideline in Appendix D.

The annotation process is completed in two turns. First, a post-doc researcher who works on discourse relations annotated all the intra-tweet explicit relations (i.e., the connective and both arguments contributed by the same user in the same tweet). Then, we extended the scope of the annotations to include inter-tweet (i.e., an argument in a tweet relates to another argument in a different tweet, typically posted by a different speaker) and non-explicit relations. Further annotations were added by an undergraduate linguistics student collaborating with the project team. The annotation process has been supervised and coordinated by the author of this dissertation. For the annotations, we used the PDTB annotator tool[4]. The explicit and implicit relations have been annotated with the PDTB Annotator v4.6 until January 2020. The hypophora relations are annotated with v4.9. We describe the annotation scheme and the quality assurance procedures in this section below.

7.2.1.1 Annotation Principles

We annotate all explicit connectives; in case of implicit relations the connective is chosen by the annotator (with the help of a list of possible connectives). For each relation, the two arguments are identified and the connective sense is labelled according to the PDTB 3.0 relational taxonomy in Figure 7.1. We primarily used the list of 100 explicit connectives from the PDTB corpus [Prasad et al., 2008] to identify connectives. Additionally, we found a few new connectives in our corpus such as *by the way, plus, so long as*, and *when-then.* If we annotated an ambiguous connective with more than one relational reading, we assigned multiple senses to the connective. For hypophora relations, only the argument spans are labelled (Arg1 for the question, Arg2 for the response). We only annotated *Arg1*, *Arg2* and the explicit or implicit connectives (and nothing else, such as attribution features/source or supplementary spans). Up to two senses were annotated if necessary for implicit and explicit relations.

Relations and their connectives can occur in intra-tweet (see example 7.9) or inter-tweet contexts (across tweets) (see example 7.10). Some authors may compose their messages in more than one tweet, i.e., they begin with one tweet and continue writing in the subsequent (adjacent) tweets, or a user may connect his own message with a previous (adjacent) message by another author.

(7.9) *Black folks in Alabama organized.* And **WON!** [Single Tweet]

(7.10) *Like I said, you don't know the whole situation to make such a judgement.* [Tweet1]
And **until you have raised one yourself, sit down and shut up!** [Tweet2]

Twitter texts, like spoken conversations, most often represent spontaneous use of language, and thus contain instances of fragmented or incomplete utterances. In our annotation, we often encounter constructions that comprise only nouns or noun phrases, but nevertheless, are often seen to stand for a complete proposition. These phrases can function as the arguments of a connective. (e.g., "*NO PROB* BUT **WHERE THE HELL DID U**", or "If **he could work on that,** *good prospect.*"). We accordingly use more flexible argument selection criteria in order to accommodate such (elliptical) structures, in addition to clauses and other constructions (nominalizations, VP-conjuncts, etc.) that typically constitute arguments in the PDTB. Furthermore, similar to the genre of instant messaging, the Twitter texts contain a wide range of acronyms for sentences/clauses that act like fixed expressions. Examples include: "idc" = I don't care; "idk" = I don't know; "idrk" = I don't really know. In our annotation, we pay special attention to these acronyms as to whether they constitute (part of) the argument of a connective or not. For example, *idc* in "idc *if* u do or not" is annotated as an argument (of *if*), while *idk* is not part of the argument in "i get your point... *but* idk the k-exol who he was talking to was comforted...".

7.2.1.2 Quality Assurance

1. **Inter-Annotator Agreement** As previously stated, this corpus is annotated in two steps. The initial stage is to annotate the explicit intra-tweet annotations, followed

[4] https://drive.google.com/file/d/1b3n7CDLoT1bPxkp5lHLC_kEHVCUNsaEk/view

by the addition of the remaining annotations in the second step. For the first and second steps, we conducted inter-annotator agreement (IAA) studies separately.

IAA Study (First round): After the annotation of explicit intra-tweet relations completed, we conducted an Inter-Annotator Agreement study on 50 threads that are randomly selected. This sub-corpus consists of 683 tweets whose average length is 188 characters, and was re-annotated by a research assistant. We calculated the percent agreement for connective detection (i.e., the percentage of connectives marked by both of the annotators), Arg1 and Arg2 span selection, and all levels of sense assignment. Arg1, Arg2 and sense agreements are calculated for the relations annotated by both of the annotators. Table 7.1 shows the percent agreement for *exact match* and *partial match* of the selected text spans. We consider one character difference in the begin & end indices of text spans as an instance of *exact match* to eliminate disagreements because of the involvement of punctuation at the end or beginning of the texts in marked spans. In *partial match* statistics, in addition to exact matches, the argument spans having any overlapping tokens are also considered matching. We manually inspected all cases of *partial match* and observed that in all cases, one annotator's argument span is fully included in the other annotator's span.

Type	% Exact	%Partial
Connective Detection	70%	-
Arg1 Span	62%	90%
Arg2 Span	89%	92%

Table 7.1: IAA for text spans (First round)

The agreement was generally good for the explicit intra-tweet relations, except for exact argument spans for Arg1. The main reason for this is the difficulty in Twitter to determine utterance and clause breaks. There was major disagreement with respect to social media specific items like hashtags and emoji (e.g., Should they be included in the argument span or not?). Social media text is genuinely more difficult to annotate than news text in this regard, and we adapted the annotation guidelines accordingly (see in Appendix D) to develop clear instructions for these cases in the second round of annotations.

Table 7.2 shows IAA statistics for sense levels[5]. Although the agreement statistics decrease as the level of the sense increases, the annotators still had a good agreement on the sense levels.

Sense Level	%
Level-1	88%
Level-2	82%
Level-3	76%

Table 7.2: IAA for sense annotations (First round)

IAA Study (Second round): After the annotation of explicit inter-tweet relations and non-explicit (i.e., implicit and hypophora) relations are added to the

[5] Level-1 specifies four sense classes, TEMPORAL, CONTINGENCY, COMPARISON, and EXPANSION. Level-2 provides 17 sense types, whereas Level-3 encodes only the *directionality* of the sense in the PDTB-3 schema (e.g., REASON vs RESULT as subtypes of the Level-2 sense type CAUSE).

corpus, we conducted another IAA study mainly for checking the agreement for implicit relation annotations, on 20 randomly selected threads. This sub-corpus consists of 267 tweets whose average length is 187 characters, and was re-annotated by a research assistant for IAA computation. We calculated the percent agreement for Arg1 and Arg2 span selection (for explicit relations), and relation sense (for implicit relations). As shown in Table 7.3, agreements on the argument spans are quite high for the explicit relations. The results are better than the first IAA study, which is likely due to the addition of less ambiguous span guidelines for social media symbols like hashtags, links, and emoticons.

Type	% Exact	%Partial
Connective Detection	71%	-
Arg1 Span	79%	93%
Arg2 Span	95%	97%

Table 7.3: IAA for explicit relation text spans (Second round)

The implicit relations annotated by both annotators were examined in this IAA study. We considered an implicit relation to be common between two annotators if an implicit relation exists in both annotations with the exact same argument spans. As a result, Arg1 and Arg2 spans of the analyzed implicit relations are always matching. Therefore, we only evaluated the sense assignments for the implicit relations. The first annotator marked 169 implicit relations, while the second annotator identified 126 implicit relations with the same argument spans as the first annotator. As a result, only these 126 common implicit relations were examined in the agreement analysis.

Sense Level	%
Level-1	68%
Level-2	45%
Level-3	41%

Table 7.4: IAA for sense annotations in Implicit relations (Second round)

Inter-annotator agreement for implicit relations is usually lower than for explicit relations, as previously discussed in the literature [Prasad et al., 2008, Zeyrek and Kurfalı, 2017, Zikánová et al., 2019, Hoek et al., 2021]. The gap between sense levels in Table 7.2 and 7.4 is rather large in our instance. We reviewed the situations of implicit relations disagreements in biweekly meetings and asked the annotator who created the annotations to be used in the IAA study to review his annotations again in light of the collective judgments. However, we were only able to reach agreement in Table 7.4. As a result, the conclusions of [Prasad et al., 2008, Zikánová et al., 2019, Hoek et al., 2021], are confirmed by these results. We also contend that implicit relation annotation is more challenging in Twitter conversations than in other written genres, owing to the ambiguous nature of texts resulting from the Twitter interface's character limit (i.e., it was 280 at the time we collected data) and spoken-like characteristics of tweets. We consider our annotations to be a silver dataset rather than gold data in terms of annotation quality because the validity of implicit relation guidelines could not be proved due to the poor agreement in the implicit relation senses.

We show the confusion matrix for the disagreements in the implicit relation sense annotations in Table 7.5. As the table shows, the annotators have significant disagreements on both high-level class annotations (e.g., Contingency vs Expansion) and finer-grained sense assignments for the same class (e.g., Expansion.Conjunction vs Expansion.Level-of-detail.Arg2-as-detail).

Ann1	Ann2	Percentage
Comparison.Concession.Arg2-as-denier	Expansion.Conjunction	4.2%
Comparison.Contrast	Comparison.Concession.Arg2-as-denier	4.2%
	Expansion.Conjunction	2.8%
Contingency.Cause.Reason	Expansion.Equivalence	4.2%
	Expansion.Level-of-detail.Arg2-as-detail	4.2%
	Expansion.Conjunction	2.8%
Contingency.Cause.Result	Expansion.Conjunction	6.9%
	Expansion.Level-of-detail.Arg2-as-detail	1.4%
Contingency.Cause+Belief.Reason+Belief	Contingency.Cause.Reason	4.2%
Contingency.Cause+Belief.Result+Belief	Contingency.Cause.Result	2.8%
Expansion.Conjunction	Expansion.Level-of-detail.Arg2-as-detail	5.6%
	Contingency.Cause.Result	2.8%
Expansion.Substitution.Arg2-as-subst	Expansion.Conjunction	4.2%
	Comparison.Contrast	2.8%
	Contingency.Cause.Result	2.8%
Temporal.Synchronous	Contingency.Cause.Result	1.4%
	Contingency.Cause.Reason	1.4%
	Expansion.Conjunction	1.4%

Table 7.5: Confusion matrix for differences in sense assignments between annotators for implicit relations

2. **Review of Annotations** All the annotations are reviewed by a trained linguistics student. He double-checked the allocated sense levels and marked annotation spans for all relations. If he disagreed with any of the markings, he presented the cases in the weekly project meetings, where they were discussed with the author of this thesis and two project leaders. The resolutions are represented in the corpus after the discussions. The reviewer also examined whether the annotations adhered to the specifications outlined in the guideline in Appendix D, and revised those that did not.

7.2.2 Quantitative Analysis

TwiConv Coherence Relations

We annotated 2281 relations in total of which 1433 are explicit relations, 732 are implicit relations and the remaining 116 are hypophora relations. Relative distribution of implicit and explicit relations is shown in Figure 7.2 for each thread. As expected, the graph indicates that there is a positive correlation between the number of explicit and implicit relations.

We observe that explicit discourse relations are a frequent occurrence in our Twitter data. Out of 1756 tweets, 47% contain at least one and 22% contain more than one discourse connective. Distribution of discourse connectives across tweets is shown in Figure 7.3. Tweets include zero to six discourse connectives, as illustrated in the bar chart. Tweet samples with 1-, 2-, 3- and 6-connectives are given in examples 7.11, 7.12, 7.13, and 7.14, respectively.

(7.11) #ketodiet update: I have been on my keto diet for a week &[6] have lost 7.60lbs.

[6] An orthographical variant of “and”

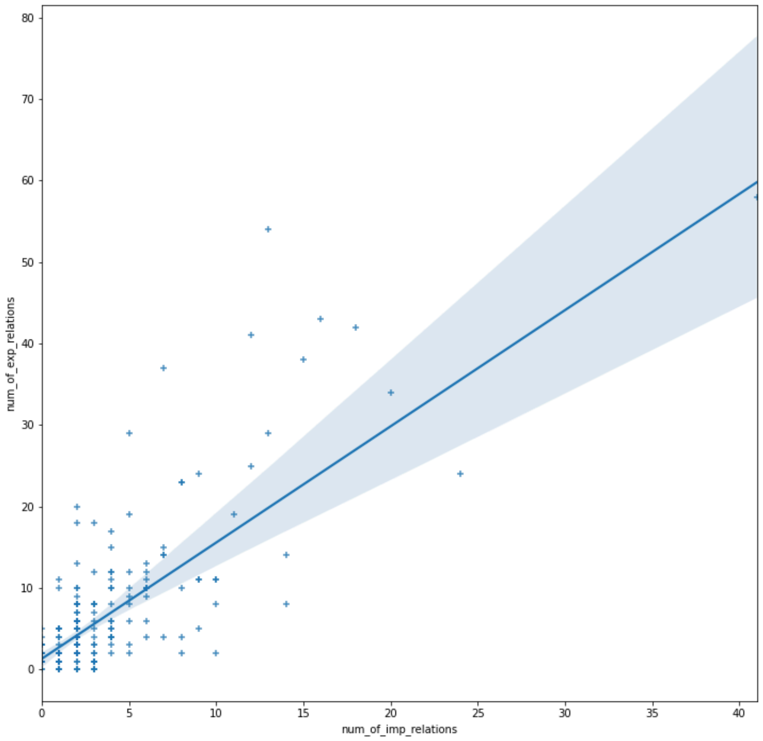


Figure 7.2: Distribution of Implicit and Explicit Relations for each thread

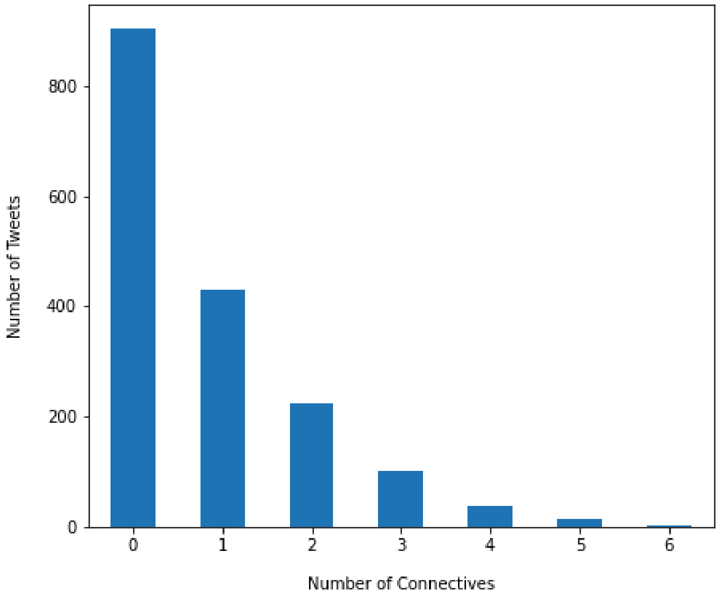


Figure 7.3: Connective distributions in tweets

Guys, I made it a week without quitting. This is a huge accomplishment. (we all recall the whole 30 diet I did for 3 days in 2016 lolz) 1 week down, many more to go! https://t.co/KCOq3C25qL

(7.12) The investigation is not over bc[7] Mueller is following the money, a process that took over two years for Nixon. Collusion with a foreign adversary to influence US electronics is treason which is against the law, so it depends on the goal of collusion. Flynn is indicted for lying

(7.13) The only Russia collusion occurred when @HillaryClinton conspired to sell US Uranium to a Russian oligarch while she was #SecretaryofState. #RussiaCollusion #UraniumOne #ClintonFoundation Why is the mainstream media so quiet? Probably because it doesn't fit their narrative.

(7.14) Yes, but if it were true and she has decided to run in 2020, it gives more people something to rally behind, a reason to get out and vote this year, a Democratic Congress when she arrives! I'm all in, and think an Oprah run would greatly help in 2018 Mid Terms! #Oprah2020

Table 7.6 shows the distribution of intra- and inter-tweet relations. As seen in the table, the majority of Explicit and Implicit relations occur within a single tweet, whereas Hypophora relations are typically inter-tweet relations. 98.5% of the inter-tweet relations span into two tweets, as illustrated in 7.15 and 7.16 for an implicit relation and an hypophora relation, respectively, but there also exist relation instances that span into three tweets (1.5%), as illustrated in 7.17. Inter-tweet relations typically occur between the tweets posted by different users (81%) but they also exist between tweets posted by the same user (19%) as illustrated in 7.17 in which the tweets are part of a thread created by the same user.

(7.15) Tweet1: [..] *Time is short!!!*
Tweet2: **Not as short as your career highlights.** [..]

(7.16) Tweet1: *Higher than a the office of a Governor?? Or he's talking of the offices when turned upside down?*
Tweet2: **A speaker is higher than the governor**

(7.17) Tweet1: *Is an individualist someone who believes that her society is individualist (either because the mechanism is choice-aware, or because the structure leads to egoism)?* 9/
Tweet2: *Or is she someone who wants the society to be individualist?* 10/
Tweet3: So, **we could imagine having someone who lives in a society where everyone is altruistic, but the society itself does not penalize for egoism, while wishing to live egoistically, but failing to due to his mis-estimation of penalties.** 11/

Relation Type	intra-tweet	inter-tweet
Explicit	90%	10%
Implicit	88%	12%
Hypophora	4%	96%

Table 7.6: Intra- and inter-tweet relation distributions

[7] An orthographical variant of "because"

TwiConv vs PDTB

Table 7.7 shows the average lengths of relation arguments for both TwiConv and PDTB[8]. The average argument lengths in the PDTB corpus are longer than those in the TwiConv corpus, which is to be expected given the character limit in Twitter posts.[9] For both corpora, the argument lengths in implicit relations are longer than those in explicit relations. For both implicit and explicit relations in the TwiConv corpus, Arg1 is on average longer than Arg2. The largest difference between Arg1 and Arg2 lengths is observed for Hypophora relations for both of the corpora, with Arg2 spans being considerably longer on average than Arg1 spans.

Corpus	Relation Type	Arg1 length	Arg2 Length
TwiConv	All	50.4	47
PDTB	All	86.9	83.3
TwiConv	Explicit	47.1	42.7
PDTB	Explicit	77	70.8
TwiConv	Implicit	56.6	50.4
PDTB	Implicit	88.7	91.5
TwiConv	Hypophora	51.4	76.7
PDTB	Hypophora	55.8	91.7

Table 7.7: Average argument lengths for TwiConv and PDTB Corpora

Twitter posts in our data, although they are written, are part of interactive conversations. While annotating these posts, we encountered a number of phenomena that are closer to spoken language than written language. Crible and Cuenca [2017] argue that discourse markers in spoken genres are more multi-functional than in written genres. In case of connectives, this indicates more diversity in spoken genres in the sense distributions of discourse relations established by a particular connective. The most frequent connectives in our corpus and in PDTB corpus are shown in Tables 7.8 and 7.9[10], respectively[11]. The connectives *and* and *but* are the most frequent connectives both in our Twitter conversations and in PDTB news data. A comparison of the highest level of sense types (i.e., class in PDTB terminology) for the explicit relations established by these top two connectives (shown in Table 7.10) reveals that "and" is used to establish TEMPORAL relations (as illustrated in Example 7.18) in 8.2% of explicit relations in TwiConv, but it is not used for that purpose in PDTB. Tonelli et al. [2010] also observed a similar pattern in their spoken dialog annotation in Italian that the connective "e" (i.e., an equivalent of "and") can express TEMPORAL relations, as well as the EXPANSION relations. Furthermore, in TwiConv, the COMPARISON relations established by "and" are substantially more common than in PDTB (5.7% vs 0.03% , respectively). The major function of "but" appears to be to build COMPARISON relations in both corpora. As a result, we do not see the same variation for "but". However, while "but" does not show the same functional variation, the difference in the usage of "and" is noticeable, leading to the idea that TwiConv is more comparable to spoken language in terms of connective functionality than written language. However, a more in-depth investigation is needed to support that claim. In addition, we rarely (if ever) annotated connectives like (if any) *since, therefore* or *nev-*

[8] We generated the PDTB statistics directly from the PDTB 3.0 corpus by using our scripts.

[9] Character limit was 280 at the time of data collection.

[10] The calculated percentages are case insensitive (e.g., "and" and "And" are considered as different instances of the same connective).

[11] Our Twitter data also exhibits different spellings for the same connective, for example, 'wen' = *when*; 'cos', 'cus', 'cuz' = *because*; 'btw' = *by the way*; '&', '&amp;', 'an' = *and.* We considered these alternative forms as orthographical variants of the same connective.

ertheless, which are typically found in formal writing (e.g., newspaper, scientific genre) [Crible and Cuenca, 2017].

Connective	Percentage
and	27.6%
but	15.9%
if	7.9%
so	6.6%
when	6.2%
because	5.7%
or	2.8%
also	2.8%
as	2.2%
then	1.8%

Table 7.8: Top ten connectives in the TwiConv explicit relations

Connective	Percentage
and	26.3%
but	15.2%
also	7.1%
if	4.7%
when	4.3%
while	3.3%
as	3.3%
because	3.1%
after	2.1%
however	2%

Table 7.9: Top ten connectives in the PDTB explicit relations

Connective	Corpus	COMPARISON	CONTINGENCY	EXPANSION	TEMPORAL
and	TwiConv	5.7%	4.0%	82.2%	8.2%
and	PDTB	0.3%	2.7%	97%	-
but	TwiConv	97.8%	-	2.2%	-
but	PDTB	98.7%	0.1%	1.2%	-

Table 7.10: Level-1 sense distributions for "and" and "but" (case insensitive)

(7.18) [..] *I'm going to create a totally new arbitrary number* <u>and</u> **assign meaning to it**.

Table 7.11 depicts the relative frequency distribution of class labels in our corpus and in the PDTB corpus. The tables show that our Twitter data has a lot more CONTINGENCY relations than the PDTB. Inline with this observation, connectives expressing CONTINGENCY relations like *if*, *when*, *because* and *so* occur relatively more frequently on Twitter as shown in Tables 7.8 and 7.9. When we are annotating the conversations in our data, we observed that longer threads frequently represent argumentative discussions. The above-mentioned CONTINGENCY expressing connectives might be considered as an evidence for an argumentative nature of conversations (i.e., users are substantiating their arguments in an argumentative context). In comparison to social media conversations, newspaper writing appears to favor narrative (TEMPORAL) and EXPANSION relations.

In terms of the distinctions between implicit and explicit relations, the table shows that in the TwiConv corpus, CONTINGENCY relations are more common in Implicit relations, whereas TEMPORAL relations are more common in Explicit relations which is consistent with the pattern in PDTB. In PDTB, Implicit relations contain a considerably lower number of COMPARISON relations than the Explicit relations (11% vs 25%, respectively) whereas in the TwiConv data, both relation types contain a similar number of COMPARISON relations (23% vs 25%).

7.3 Outlook: Interplay of Coherence Relations and Coreference

The contribution of the work in this chapter is to enrich TwiConv with coherence relations, while subsequent analyses are out of scope. In this section, we present various studies

Class	Relation Type	% in Twitter	% in PDTB
EXPANSION	All	32%	44%
EXPANSION	Explicit	33%	42%
EXPANSION	Implicit	30%	46%
CONTINGENCY	All	34%	25%
CONTINGENCY	Explicit	29%	16%
CONTINGENCY	Implicit	43%	35%
COMPARISON	All	24%	18%
COMPARISON	Explicit	25%	25%
COMPARISON	Implicit	23%	11%
TEMPORAL	All	10%	13%
TEMPORAL	Explicit	13%	17%
TEMPORAL	Implicit	4%	8%

Table 7.11: Average argument lengths for TwiConv and PDTB Corpora

from the research areas that could potentially exploit the coherence and coreference data together. Our goal is not to provide a comprehensive review of the literature in these areas, but rather to provide an overview for potential future research using our dataset by highlighting pioneer and/or contemporary studies in the field.

There exist three main lines of research that exploit the type of data that we provide. The first line focuses on the impact of coherence relations on referential accessibility, both theoretically and empirically. The second line of research, in a related direction with the first one, investigates the impact of knowledge on coherence relations in automatic coreference resolution. Finally, the third line is interested in the effect of coreferential information on the identification of coherence relations.

7.3.1 Referential accessibility

Theories on discourse relations have indications on the accessibility of the referents. For instance, SDRT adapts the right frontier constraint [Polanyi, 1988, Webber, 1988] for addressing the accessibility of the elements in a discourse which can serve for an antecedent for anaphoric constructions such as ellipsis and pronominal expressions [Asher and Lascarides, 2003]. Potential antecedents of an anaphor that are placed at the right frontier of a discourse unit, according to the framework, can be accessible more easily than antecedents placed elsewhere. Holler and Irmen [2007] makes an empirical validation of this constraint for resolving inter-sentential anaphora through a questionnaire survey. The experimental setup includes items for antecedent and anaphor pairs of different genders and same genders. Their findings indicate that patterns obtained for same gender pairs point the Right Frontier effect.

Fox [1987] makes a corpus study where the discourse structure is represented by RST trees. In her study, Fox [1987] examines the use of anaphoric third person singular human references (pronouns and noun phrases) in spoken (face-to-face and telephone conversations) and written texts (newspaper articles and psychoanalytic biography). Fox argues that linear anaphoric distance (i.e., textual distance between pronouns and their antecedents) is not enough for explaining the form of referential choice (e.g., pronoun or full NP) whereas investigation of discourse structural features can give more insight about the referential choice and hence, more insight about accessibility of the antecedents. Fox argues that for the explanation of the characteristics of the long-distance pronominalization (i.e., a referent is mentioned by means of a low-information anaphoric device, a pronoun in her case, after some absence in the text), discourse structure is a key component and big portion of the data can only be explained by structural means using the RST trees. In the investigated cases, she shows that the pronouns do not establish connections with the immediately preceding prepositions. They rather connect the current proposition on the upper level of the discourse structure (i.e., a tree in case of RST). As a result, she argues that although those connected propositions are not close sequentially, they are closer discourse-structurally.

Chiarcos and Krasavina [2005] conduct a corpus study and compare the rhetorical distance (e.g., length of the path between two elements in an RST tree) and the anaphoric distance as predictors of referential choice. They report that for the referential forms, referential distance is a better predictor of pronominal forms for short-distance pronominalication (i.e., the antecedent is within the last two clauses). If the referential distance is greater than 2 clauses, rhetorical distance is a better indicator. In this regard, their study supports the claims of Fox [1987].

7.3.2 Exploiting coherence relations for the coreference resolution

Cristea et al. [1999] focus on determining whether a discourse structure theory (RST in their case) can be helpful for improving the performance of an anaphora resolution system. At the time of the study, anaphora resolution systems (e.g., [Lappin and Leass, 1994]) have primarily consisted of three modules: determining potential antecedents (COLLECT), eliminating potential antecedents that are incompatible with the anaphor (FILTER), and selecting the best fitting antecedent based on an ordering policy (PREFERENCE). Cristea et al. [1999] hypothesize that the size of the search space required to resolve referential expressions differs significantly depending on whether a linear (looking at the discourse units that immediately precede the anaphor) or discourse (looking at the discourse units are hierarchically preceding the anaphor the anaphor) model is used. If a discourse model is more efficient than a linear model, it is supposed to provide a smaller search space for the potential antecedents because an anaphora resolution system, then, will have to consider fewer options which will reduce the probability of errors. They make an empirical comparison of potential linear and hierarchical discourse models to resolve the coreferential links in the texts. The units that are hierarchically preceding a given unit are determined by following the implications of Veins Theory (VT) [Fox, 1987], indicating that the distinction between nuclei and satellites in RST constraints the range of referents to which anaphors can be resolved. Newspaper texts that are manually annotated for coreference relations and rhetorical structure trees are used for the analysis. The analysis in the paper suggests that by taking advantage of the hierarchical structure of texts, an automatic system can increase the potential for accurately identifying coreferential links. If all the discourse units in the preceding discourse are considered equally, the increase is only statistically significant when a discourse-based coreference system looks at a maximum of four discourse units back to establish coreferential links. However, assuming that proximity is important in the formation of coreferential links and that referential expressions are more likely to be linked to referents who have recently been used in discourse, the increase is statistically significant regardless of how many units a discourse-based coreference system looks back.

This line of research did not receive much attention since from Cristea et al. [1999]. Among the few studies, Khosla et al. [2021] is interesting as they incorporated the discourse structural features into a neural coreference system and obtained statistically significant improvement in the system performance. The OntoNotes corpus [Pradhan et al., 2012] which contains multiple sub-genres including news articles and telephone conversations and RST subset of the ARRAU corpus [Poesio et al., 2018] are used in the experiments.

They first converted the documents into hierarchical RST trees by using an automatic RST parser [Yu et al., 2018] and then introduced discourse level constraints extracted from these RST trees into a neural baseline system [Lee et al., 2017]. Discourse level features are composed of distance- and coverage-based features that are computed based on the placement of the lowest level node in the tree (referred to as LCA in the study) covering a mention pair of the current mention and a candidate mention. The distance-based feature, for example, considers the distance between a mention and the root of the LCA in the tree covering that mention and a candidate mention. The lower this distance, the closer the two mentions in the discourse are regarded to be. Incorporating these features resulted in a minor (*approx* 1%) but statistically significant improvement in baseline performance, indicating that discourse coherence information could be valuable for coreference resolution.

7.3.3 Exploiting coreference features for the identification of coherence relations

Rutherford and Xue [2014] built a statistical classifier to investigate the impact of Brown clusters (i.e., clusters of words that are similar) and coreferential features for inferring the implicit relations in PDTB corpus [Prasad et al., 2008]. They consider various coreference patterns as features such as the number of shared coreferential pairs between two arguments and a binary feature representing whether the main verbs of the two arguments have the same subjects or not (i.e., the subjects are coreferential or belong to the same Brown cluster). The coreference chains are resolved automatically through Stanford CoreNLP [Lee et al., 2013]. Their experiments show that Brown clusters and coreference features are effective for improving the performance for all type of discourse relations. The ablation study indicates that coreference features are more distinctive for CONTINGENCY and TEMPORARY relations (decrease in F1 scores when coreference features are eliminated from the feature set is $\approx$ 1%) than EXPANSION and COMPARISON relations. Rutherford and Xue [2014] also make a quantitative investigation of coreferential patterns in discourse relations. TEMPORAL relations have a greater coreferential rate (i.e., the number of coreferential pairs between relation arguments) than the other relations, according to their findings. In addition, CONTINGENCY relations have the lowest coreferential rate when they only look at subject-coreferential rates.

Dai and Huang [2019] incorporate external event knowledge and coreference information into neural network models for discourse parsing. For event knowledge, common sense knowledge (e.g., smoking causes the lung-cancer) is extracted from different resources such as ConceptNet [Speer and Havasi, 2012]. They generate the coreference chains in the text by the Stanford coreference resolver [Clark and Manning, 2016]. They integrate the event knowledge and coreference information to their implementation of a neural baseline system through a separate knowledge layer which also contains knowledge regularization component to evaluate the applicability of the integrated features in a particular context. They experimented in several setups with Glove [Pennington et al., 2014] and ELMo [Peters et al., 2018] Word Embeddings. They evaluate the model on PDTB-2 [Prasad et al., 2008]. The experiments indicate that their approach improves the performance of both implicit and explicit discourse relation recognition compared to previous work. Qualitative analysis indicate that by incorporating coreference relations into the base model using the regularization approach, implicit discourse relation prediction performance was improved for two classes, Contingency and Expansion. They also argue that the model without knowledge regularizer performs significantly worse than the base model indicating that using external knowledge or linguistic constraints blindly can decrease the performance. Therefore, they conclude that the knowledge regularizer plays a key role in achieving the state-of-the-art performance and the knowledge layer must be used together with knowledge regularization in their framework. Their best result is achieved with ELMo embeddings and overperforms the previous best result [Bai and Zhao, 2018] with 1.8% F1 score.

In a more recent study, Guz and Carenini [2020] investigate the benefits of coreference information for neural RST discourse parsing. They first implement an RST parser as a baseline using the SpanBERT contextualized word embeddings [Joshi et al., 2020]. They applied two models to investigate the effect of coreference information on discourse parsing. For the first model, the output of Joshi et al. [2020]'s SOTA coreference resolver is given as input to the implemented RST parser in a similar way to Dai and Huang [2019]. The second model using the coreference information is a more sophisticated neural model where the discourse parsing and coreference resolution are performed simultaneously in a multitask learning setup based on the sharing of SpanBERT encoder. Their models outperform all previous approaches, indicating strong benefit of pre-trained language models for RST discourse parsing. However, they do not observe statistically significant improvement in performance of the baseline when utilizing the coreference information. They state that this result can be explained in different ways. For instance, it is possible that coreference information is not a strong signal for discourse structure or the automatically extracted coreference information can be too noisy. They note that further studies are needed to explore these possibilities.

7.4 Summary

We presented a description of our PDTB-style annotation of English Twitter conversations. Given that Twitter conversations are a genre with the discourse structure of written multi-party interactions, they provide an appealing material for this type of annotation. Our initial analysis of the annotations revealed that despite being written, Twitter conversations show properties typical of spoken interactions.

We believe that this data could be relevant in conjunction with the research provided in this dissertation in two directions:

1. In Chapter 4, we provided a corpus research in which we compared different genres using quantitative aspects of anaphoric distance in terms of sentences, tokens, and NPs. The shallow discourse relations described in this chapter can be used to broaden the scope of the quantitative comparison carried out in Chapter 4.

2. We investigated the problem of adapting coreference resolution systems to TwiConv data in Chapter 6. Coherence relations would be beneficial for improving the performance of coreference resolution techniques, according to the studies presented in Section7.3.2. Therefore, we believe that this data is potentially helpful for the adaptation of coreference resolution to the domain of Twitter conversations.

The majority of the studies presented in 7.3.3 indicates that coreferential links are potentially useful for the identification of coherence relations. One counter example is the study of Guz and Carenini [2020] where the authors argue that noisy coreference links generated by the automatic systems might not be robust enough for helping the improvement of the baseline systems. As we discussed in Chapter 6.4.2, performance of existing coreference resolution systems decrease sharply when run on conversational Twitter data. Therefore, it can be assumed that an automatic coreference resolution tool run on this data will produce more noisy results than the written texts. We believe existing gold coreference annotations on this data will be very beneficial for the research investigating the interplay if coherence and coreference for social media texts in general.

8

Conclusion

This concluding chapter first provides a summary of the contents of the previous chapters, which covers an overview of the main contributions and key findings of this dissertation. An outlook on future work is then presented in Section 8.2, which outlines the limitations of our work, and explores potential directions to resolve them and to broaden the overall scope of this study.

8.1 Summary

In this thesis, we explore how the language used in spoken and written modes, and the text genres generated in these modes relate to each other in terms of nominal coreference establishment strategies. We are particularly interested in a genre of texts that is formed of conversational interactions on the Twitter platform. We investigate the linguistic and computational aspects of nominal coreference in these Twitter conversations and we determine how this genre relates to other text genres in spoken and written modes.

We employ empirical methods for addressing our research questions. We apply corpus-based, psycholinguistic and computational approaches, which we believe complement each other.

To the best of our knowledge, Twitter conversations have not yet been analyzed for coreference in the literature. One factor for the lack of research on this genre may be the unavailability of coreference-annotated corpora composed of Twitter conversations. To address this shortcoming, **we first built a new corpus of such conversations which we call TwiConv**. We collected English tweets using the Twitter API and constructed the conversation structures by tracking the reply_to_id's that is present in tweet information. We then manually annotated the TwiConv corpus for coreference relations. For this purpose, we adapted a set of existing coreference annotation guidelines according to the characteristics of Twitter data. In order to validate the guideline and the annotation scheme, we conducted an inter-annotator-agreement study, and the results showed that the annotators had high levels of agreement (Krippendorff's $\alpha \geq .800$). We also included automated checks to guarantee the robustness of the annotations. Our observations during the annotation process indicate that the texts in the Twitter conversations had unusual referential expression forms. For instance, non-standard nominal forms such as usernames and hashtags as well as emojis are frequently used as referring expressions. Additionally, as the platform enables multi-modal communication by allowing the integration of visual materials and external URLs into the texts, exophoric references to the entities in the visual media also exist in the texts. Furthermore, Twitter allows for multi-user interactions, which leads to additional challenges. For instance, in a Twitter conversation, first, second, and third person pronouns may all refer to the same entity, which usually does not happen in dialogues and monologues. **These and many other characteristics**

make the genre of Twitter conversations challenging for automatic coreference resolution systems.

Chapter 4 presents an in-depth comparative corpus-based study that examines the widely used coreference-annotated corpora OntoNotes and Switchboard, as well as the TwiConv corpus that we built from scratch. Since we compare texts from different resources, we first harmonized the datasets to make them comparable in terms of technical format and annotations. This is, as far as we are aware, the first study comparing the patterns in OntoNotes and Switchboard. Altogether, the three corpora include texts from a range of spoken and written genres, namely telephone conversations (both in Switchboard and OntoNotes), broadcast conversations, broadcast news, newswire, web blogs and Twitter conversations. We use linguistically motivated features of nominal coreference and conduct a quantitative comparison of genres. Our analyses collectively indicate that there is a ranking of the genres, which suggests a continuum for the language used in spoken and written modes. For the genres near the extremes of the continuum (e.g., telephone conversations and news), differences in the numeric values of the analyzed features are more pronounced, which clearly indicates that **text genre has an impact on emerging coreference patterns**. The position of Twitter conversations in the continuum varies depending on the feature. However, it is mostly situated between the spoken and written genres, typically being closer to the spoken end in terms of the quantitative values of the examined features.

In a number of theoretical approaches on referential choice, textual distance between the referring expressions (i.e., anaphoric distance) is described as a key component for the assessment of the cognitive status of an entity, and as a result, it has a significant impact on referential choice Chafe [1976], Givón [1983], Ariel [1990], Gundel et al. [1993], von Heusinger and Schumacher [2019]). In addition to the other features we examine, our study in Chapter 4 suggests that anaphoric distance also varies across genres. In order to see whether the language mode alone has any effect on anaphoric distance, we carry out a crowdsourced story continuation experiment. Chapter 5 describes the experiment and presents the findings. Participants of the experiments are hired via crowdsourcing. To ensure data quality is high, we employ a number of techniques, which are presented in detail in Section 5.2. In the experiment, participants are presented with written and audio stimuli, and they are asked to continue the stories in these stimuli by providing spoken and written responses. We introduce a main character in each story, and we check the first reference to the main character in the participants' responses. The key observed metric in our study is the number of clauses between referring expressions (i.e., clause-based distance). In order to control the main character's recency and, by extension, the distance between the main character and any potential references in the responses, the lengths of the stories in the stimuli are systematically manipulated. We looked at the referential form of the participants' first references to the main characters in the stimuli and examine the differences in spoken and written responses. Only 5% of the responses had to be removed for quality-related reasons after manual review, indicating that we were successful in leveraging crowdsourcing to collect high-quality data for both spoken and written forms. The descriptive and inferential statistics we employed in the analysis indicate that spoken responses allow greater distances than written responses, suggesting that **language mode has an impact on anaphoric distance**.

Automatic Coreference Resolution (ACR) is a crucial step in many NLP applications that seek to extract information from texts. Due to the non-standard language and the unusual forms of nominal reference they contain, tweets are a difficult genre for ACR. In Chapter 6, we present a study that retrains a state-of-the-art end-to-end neural ACR system [Lee et al., 2018] with in-domain (i.e., the TwiConv corpus) and out-of-domain (i.e., different subsections of the OntoNotes corpus) data to improve its performance on conversational Twitter data. We empirically investigate whether coreference resolution on Twitter conversations would benefit from mode-based separation in external training data. To explore this research question, we combined the in-domain TwiConv data with different sections of OntoNotes that differ from each other in terms of genre and mode, and we ran a series of experiments to find the best combination. Experiments indicate that for coreference resolution on Twitter conversations, integrating external data in the training set is generally useful. However, including external spoken data is more beneficial than written data. As a result, we show that **for the out-of-domain training data, the choice of genres, and the language mode, can make a bigger difference than simply scaling the size of the training data set**. This finding suggests that the out-of-domain training data should be carefully chosen for domain adaptation research. **We improve the performance of the automatic coreference resolver on Twitter conversations by 21.6% F1 score**. We thus provide a competitive baseline for the TwiConv corpus.

We are interested in learning more about the overall discourse behavior of Twitter conversations. To address this research interest, in addition to the referential expressions and coreference chains, we also annotated coherence relations on the TwiConv data. Because the structure of Twitter conversations is under-explored in the literature, we did not apply a discourse framework that makes assumptions about the overall structure of texts, such as RST [Mann and Thompson, 1988], which models discourse as a tree structure. Instead, we applied the PDTB approach [Prasad et al., 2018], which makes no assumptions about the discourse structure and is more concerned with capturing local coherence between adjacent or textually proximate units than capturing the overall structure of the discourse. In Chapter 7, we describe the annotation procedure and present the findings of an initial analysis that compares the coherence relations between the TwiConv corpus and the PDTB corpus, which is made up of edited news texts. We observed that **despite being written, Twitter conversations show properties typical of spoken interactions**. For instance, TwiConv data is more comparable to spoken language in terms of the diversity in connective functionality than written language. We undertook an inter-annotator-agreement study to confirm the validity of the guidelines and the annotations. For the explicit relations (i.e., the relations that are signalled by explicit discourse connectives), we acquired a high agreement score (see in Table 7.3). However, after repeating the agreement study for the implicit relations iteratively over several rounds and having in-depth discussions regarding the disagreements in-between the rounds, we were still unable to reach a high level of agreement among the annotators for implicit relations. Although it has been previously reported for the PDTB-style annotations that implicit relations are more difficult to annotate than explicit relations (e.g., by Zeyrek and Kurfalı [2017], Zikánová et al. [2019]), our agreement score for implicit relations (45% agreement ratio for the sense (Level-2) in Table 7.4)) remained lower than what, for instance, Zikánová et al. [2019] reported (i.e., 57.7% agreement ratio). Therefore, we draw the conclusion that **it is more challenging to annotate implicit relations in Twitter texts than it is in edited written texts** (e.g., journalistic texts in case of Zikánová et al. [2019]). Here, we only compare two studies; future research may shed additional light on the differences across genres in the annotation of implicit relations.

8.2 Limitations and Outlook

A number of extension or improvements can be envisaged in future work. For instance, the comparative corpus analysis that is described in Chapter 4 can be expanded in a number of ways. First, we do not count null pronouns (for example, in the subject position of non-finite sentences) or bridging anaphora as instances of references in our corpus-based analysis when computing the quantitative features. However, it is possible that these phenomena will affect the metrics that were presented, particularly the distance-based metrics. We believe the distance question could be further clarified by a corpus analysis that takes these phenomena into account (which will however require substantial manual annotation).

One outcome from the corpus study in Chapter 4 is that we were unable to arrive at firm conclusions regarding the token-based anaphoric distance when comparing genres and modes (in contrast to the clearer results for clause-based distance). Conclusive results for this metric can be obtained by another controlled experiment, similar to the one described in Chapter 5, where token-based anaphoric distances are taken into account as the experimental factor.

The third potential extension of the corpus study is related to the distance-based metrics that count the number of clauses and intervening noun phrases as units. Those metrics indicate conclusive results but point in different directions. We find that clause-based anaphoric distance is longer in spoken genres, whereas NP-based anaphoric distance is longer in written genres. This observation raises interesting research questions: Is it related to the types of clauses and tokens used in the clauses? Or is there any connection with the cognitive processes experienced during speaking and writing?

As discussed in Section 2.3.2, there exist theoretical approaches that associate the linguistic characteristics of texts with dimensions of textual variation (e.g., Biber [1988], Koch and Oesterreicher [2012]). The fourth avenue we might consider for the expansion of our corpus study is to broaden the scope of the analysis, for example, using the framework proposed by Koch and Oesterreicher [2012] and comparing the genres in terms of language of "immediacy" vs "distance" by studying additional linguistic features associated to these dimensions.

In Section 5.2.5, we show the findings of a preliminary analysis indicating that referential choice in a response in our story continuation experiment is correlated with the discourse relation established between the response and the stimulus. We also discussed the potential that the type of the stimuli (e.g., narrative or not, according to Labov and Waletzky [1967]'s framework) could affect the type of relation established by the response. We therefore hypothesize that the type of the stimulus affects the form of the first reference to the main character in a response. This hypothesis needs to be tested, for instance, by examining the relevant qualitative characteristics of all the stimuli in our experiment data.

To the best of our knowledge, the interplay of coreference and coherence in the context of Twitter conversations has not yet been studied in the research literature. Insights into how they interact may shed more light on the discourse characteristics of this genre, given that we here provide a corpus which includes both coreference and coherence annotations. Additionally, this data can be used for evaluating the methods that automatically detect coherence relations and coreference links. Implicit relations are harder to annotate in Twitter conversations than in news texts, according to our experience on the annotation of coherence relations described in Section 7.2.1. The variations across these genres could be better understood if this discrepancy were investigated in greater depth.

Finally, in a general sense, we believe that incorporating linguistic knowledge into learning-based systems can enhance the performance of a coreference resolver, particularly for non-standard texts, such as Twitter conversations. This may be especially important when there is little or no training data available. We think that investigating the techniques for integrating linguistic knowledge into ACR research—that is, creating a bridge between linguistic and statistical knowledge—can lead to higher performance for domain-adaptation tasks. For this type of further investigation, the linguistic findings we give in this study, particularly in Chapter 4, may prove to be helpful.

Appendices

Appendix A

Statistical Significance Tables

Dataset pair	NP(LS) vs PRP	NP(SS) vs PRP	NP(LS) vs 3rd PRP[1]	NP(SS) vs 3rd PRP
tw-swbd	***	***	***	***
tw-tc	***	***	***	***
tw-bc	***	***	***	***
tw-bn	***	***	***	***
tw-nw	***	***	***	***
tw-wb	***	***	***	***
swbd-tc	***	***	***	***
swbd-bc	***	***	***	***
swbd-bn	***	***	***	***
swbd-nw	***	***	***	***
swbd-wb	***	***	***	***
tc-bc	***	***	***	***
tc-bn	***	***	***	***
tc-nw	***	***	***	***
tc-wb	***	***	***	***
bc-bn	***	***	***	***
bc-nw	***	***	***	***
bc-wb	***	***	***	***
bn-nw	***	***	***	***
bn-wb	***	***	n.s.	n.s.
nw-wb	***	***	***	***

Table A.1: Statistical significance of pairwise NP vs PRP proportion differences between datasets ('***'= $p<0.05$) (The significance values are computed applying the Pearson's χ^2 with post-hoc pairwise Fischer test where the correction method for multiple comparison is set to "holm".

[1] 3rd Person Pronouns

Dataset pair	Personal PRP (PRNs included)	Personal PRP (PRNs excluded)
tw-swbd	***	***
tw-tc	***	***
tw-bc	***	***
tw-bn	***	***
tw-nw	***	***
tw-wb	***	***
swbd-tc	n.s.	***
swbd-bc	***	***
swbd-bn	***	***
swbd-nw	***	***
swbd-wb	***	***
tc-bc	***	***
tc-bn	***	***
tc-nw	***	***
tc-wb	***	***
bc-bn	***	***
bc-nw	***	***
bc-wb	***	***
bn-nw	***	***
bn-wb	***	***
nw-wb	***	***

Table A.2: Statistical significance of pairwise differences in Personal PRP distributions between datasets ('***'= $p<0.05$)

Dataset pair	Wilcoxon (Large Span)	Wilcoxon (Small Span)
tw-swbd	***	***
tw-tc	n.s.	***
tw-bc	***	***
tw-bn	***	***
tw-nw	***	***
tw-wb	***	***
swbd-tc	***	***
swbd-bc	***	***
swbd-bn	***	***
swbd-nw	***	***
swbd-wb	***	***
tc-bc	***	***
tc-bn	***	***
tc-nw	***	***
tc-wb	***	***
bc-bn	***	***
bc-nw	***	***
bc-wb	***	***
bn-nw	***	***
bn-wb	***	n.s.
nw-wb	***	***
spoken-written	***	***
spoken-tw	***	***
written-tw	***	***

Table A.3: Statistical significance of pairwise NP-length differences between datasets ('***'= $p<0.05$, n.s. (non significant) = $p>0.05$)

Dataset pair	Wilcoxon (Large Span)	Wilcoxon (Small Span)
tw-swbd	***	***
tw-tc	n.s.	***
tw-bc	***	***
tw-bn	***	***
tw-nw	***	***
tw-wb	***	***
swbd-tc	***	n.s.
swbd-bc	***	***
swbd-bn	***	***
swbd-nw	***	***
swbd-wb	***	***
tc-bc	***	***
tc-bn	***	***
tc-nw	***	***
tc-wb	***	***
bc-bn	***	***
bc-nw	***	***
bc-wb	***	***
bn-nw	***	***
bn-wb	***	n.s.
nw-wb	***	***
spoken-written	***	***
spoken-tw	***	***
written-tw	***	***

Table A.4: Statistical significance of pairwise NP-height differences between datasets ('***'= $p<0.05$)

Dataset pair	Wilcoxon test (with DM)	Wilcoxon test (without DM)
tw-swbd	***	n.s.
tw-tc	n.s.	n.s.
tw-bc	n.s.	n.s.
tw-bn	***	n.s.
tw-nw	n.s.	n.s.
tw-wb	n.s.	n.s.
swbd-tc	n.s.	n.s.
swbd-bc	***	***
swbd-bn	***	***
swbd-nw	***	n.s.
swbd-wb	***	n.s.
tc-bc	***	n.s.
tc-bn	***	***
tc-nw	n.s.	n.s.
tc-wb	***	n.s.
bc-bn	***	***
bc-nw	n.s.	***
bc-wb	n.s.	n.s.
bn-nw	***	***
bn-wb	n.s.	n.s.
nw-wb	n.s.	n.s.

Table A.5: Statistical significance of pairwise token-based distance differences between datasets ('***'= $p<0.05$)

Dataset pair	Wilcoxon test (with PRN)	Wilcoxon test (without PRN)
tw-swbd	***	n.s.
tw-tc	n.s.	n.s.
tw-bc	***	***
tw-bn	***	***
tw-nw	***	***
tw-wb	***	***
swbd-tc	***	***
swbd-bc	***	***
swbd-bn	***	***
swbd-nw	***	***
swbd-wb	***	***
tc-bc	***	***
tc-bn	***	***
tc-nw	***	***
tc-wb	***	***
bc-bn	***	***
bc-nw	***	***
bc-wb	***	***
bn-nw	***	***
bn-wb	***	***
nw-wb	n.s.	n.s.
spoken-written	***	***
spoken-tw	n.s.	***
written-tw	***	***

Table A.6: Statistical significance of pairwise clause-based distance differences between datasets ('***'= p<0.05)

Dataset pair	Wilcoxon test (with PRN)	Wilcoxon test (no PRN)	Nominal Density
tw-swbd	***	n.s.	***
tw-tc	***	***	***
tw-bc	***	***	n.s.
tw-bn	***	***	***
tw-nw	***	***	***
tw-wb	***	***	***
swbd-tc	***	***	***
swbd-bc	***	***	n.s.
swbd-bn	***	***	***
swbd-nw	***	***	***
swbd-wb	***	***	***
tc-bc	***	***	n.s.
tc-bn	***	***	***
tc-nw	***	***	***
tc-wb	***	***	***
bc-bn	***	***	***
bc-nw	***	n.s.	***
bc-wb	***	***	***
bn-nw	***	***	***
bn-wb	n.s.	n.s.	***
nw-wb	n.s.	n.s.	n.s.
spoken-written	***	***	***
spoken-tw	***	***	***
written-tw	***	***	***

Table A.7: Statistical significance of pairwise NP-based distance differences and nominal densities between datasets ('***'= $p<0.05$)

Appendix B

Hierarchical Clusters for Genre Comparison

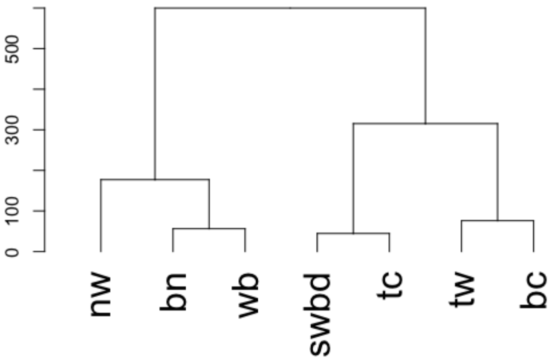


Figure B.1: Genre clustering based on NP vs PRP frequency distributions

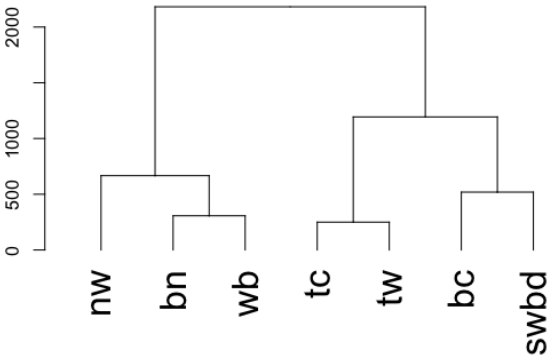


Figure B.2: Genre clustering based on NP-length

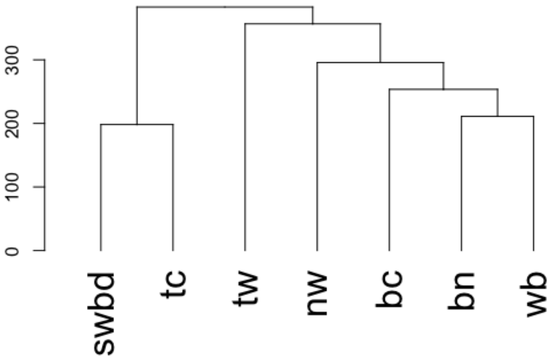


Figure B.3: Genre clustering based on TBD

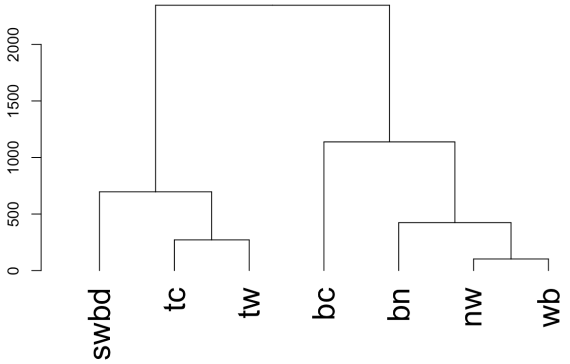


Figure B.4: Genre clustering based on CBD

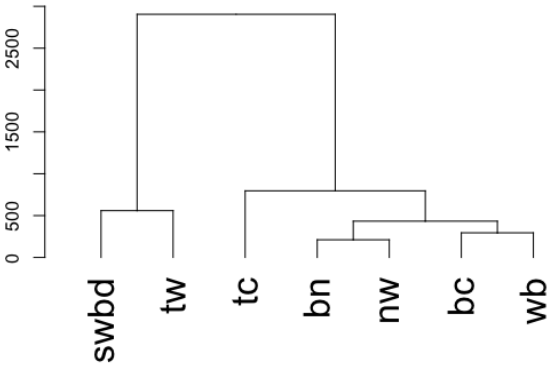


Figure B.5: Genre clustering based on NBD

Appendix C

Coreference Annotation Guideline for TwiConv

C.1 Introduction

This guideline presents the instructions for the annotation of coreference chains in conversational social media texts[1]. This version of the guideline addresses only identity coreference; non-identity reference (bridging) is not being annotated for now. For the annotation, we use freely available MMAX annotation tool[2].

In the following, Section 2 describes in detail the types of referring expressions that are subject to the annotation. Section 3 describes the annotation process, and Section 4 defines the attributes that have to be assigned to each markable.

C.2 Markables

In this section, we first discuss the various types of markables to be annotated in 2.1, and then in 2.2 provide guidance on identifying their spans. The square brackets around the expressions demonstrate the span boundaries and indices under the brackets represent the id of the coreference chain where the markable belongs to.

C.2.1 Types of markables

Syntactically, markables are phrases with nominal or pronominal heads. The following referring expressions are to be considered as markables:

1. Full nominal phrases, e.g. *a big blue sky*;
2. Proper names and titles, e.g. *Mr. Black*;
3. Pronouns
    - Personal pronouns: We annotate the personal pronouns "*I, you, my, me, mine, your, yours, they, them, their, theirs, he, him, his, himself, her, hers, her, herself, it, its, itself*".

      (C.1) [**It**]$_j$'s beginning to rain! - [Daisy]$_i$ exclaimed to [**herself**]$_i$.

      (C.2) [John]$_i$ is calling [**his**]$_i$ doctor.

[1] Most relevant portions of "Parellel coreference annotation guidelines" by Yulia Grishina and Manfred Stede (2016) are adapted to our data. The document can be found here: `https://github.com/yuliagrishina/CORBON-2017-Shared-Task/blob/master/Parallel_annotation_guidelines.pdf`

[2] http://mmax2.net/index.html

(C.3) [**She**]$_i$ is [my new lawyer]$_i$.

(C.4) [**I**]$_j$ have already payed 699 for [this]$_i$ and [**it**]$_i$ is not working.

(C.5) This is [**my**]$_i$ notebook. [**I**]$_i$ bought [**it**]$_j$ last week.

In multi-turn conversations with more than two participants, it is possible that first, second and third person pronouns refer to the same entity.

(C.6) **user1:** [I]$_i$ prefer to go to small cinemas instead of the big chains.

(C.7) **user2:** (reply-to-user1) What are [you]$_i$ talking about?

(C.8) **user3:** (reply-to-user2) [He]$_i$ is talking about supporting small business.

- Demonstrative pronouns: *this, that, these, those*

  (C.9) You need [a camera that works in the dark]$_i$. Hm, take [**this**]$_i$, [it]$_i$ has a great shutter speed.

  In the example, the demonstrative pronoun *this* corefers with the pronoun *it* in the next sentence and must be annotated.

  Predicative constructions are annotated in the following way:

  (C.10) [This]$_i$ is [a bank]$_i$, but [it]$_i$ is not very well-known.

- Relative pronouns, such as *who, whom, whose, which, that* etc.

  The relative pronouns are used to construct relative clauses in the sentence. There are different types of relative clauses for which the annotation instructions are presented below.

  If a relative pronoun is used in a restrictive relative clause, the whole NP span is annotated as one mention:

  (C.11) I met [the cyclist who won the race]$_i$. [She]$_i$ deserved that result.

  We use non-defining relative clauses to give extra information about the person or thing. In writing, commas are often put around non-defining relative clauses. In that case the modified noun and the relative pronoun are annotated separately and put in the same coreference chain.

  (C.12) I was talking about [my uncle]$_i$, [who]$_i$ has the horse, when you came.

  There exist also free (headless) relative clauses which are not used as noun modifiers, instead they serve as arguments in the main clause. In that case, we only annotate the relative pronoun.

  (C.13) I saw [what]$_i$ you cooked and ate [it]$_i$.

  Keep in mind that pronouns can be ambiguous:

  (C.14) For both India and Pakistan, Afghanistan risks turning into a new disputed territory, like [Kashmir]$_i$, [where]$_i$ the conflict has damaged both countries for more than 50 years.

  (C.15) Daisy managed to discover *where* Mr. Baccini's dishonest partner was now living and was anxiously expecting her cheque.

  In example C.14, "where" is a relative pronoun and refers to Kashmir (to confirm this, one can substitute *where* by *in which*). In contrast, in C.15, *where* is not a relative pronoun and should not be annotated.

- Question pronouns, such as *who, what, which, where* etc.

  Question pronouns are not considered as markables and not annotated.

  For instance, we don't annotate "who" in the following small conversation:

  (C.16) – **user1:** Who is calling?

  - **user2:** [Jane]$_{i}$ is on the phone, I think [she]$_{i}$ wants to visit us.

- “HIS/HERS”, “HIS or HERS” and similar forms are annotated as a single markable.

4. NPs with quantifiers

   Be careful when annotating NPs with quantifiers, e.g. *all people, two people, 105 Million euro* etc. If you are not sure about the definiteness of an NP, apply the following test: try inserting a definite article or a demonstrative pronoun. If the meaning of the phrase is not changed, then the NP is definite. Example: “*all people*” > “*all these people*” > definite NP.

   Quantifiers (of the form “X of Y”) should not be coreferenced with the entities they modify:

   (C.17) “[a mile of [highway]$_{j}$]$_{i}$”

   (C.18) “[a group of [doctors]$_{k}$]$_{l}$”

   Similar constructions such as “both of those things” and “all of my friends”, the markable spans are similar but coreferential strategy is different: Since “my friends” and “all of my friends” would usually be equivalent/refer to the same group of people, both (“my friends” and the full phrase “all of my friends”) are selected as markables and they are marked as coreferent.

5. Nominal premodifiers

   Nominal premodifiers are annotated as separate markables only if there is an overt reference to that modifier in the text as in C.19. Otherwise they are not annotated separately and included in the span of the NP they modify (C.20).

   (C.19) I bought a new [[ceramic]$_{i}$ pot]$_{j}$. I really like [that material]$_{i}$ for cooking because nothing sticks on [it]$_{i}$.

   (C.20) I bought a new [ceramic pot]$_{i}$. [It]$_{i}$ is perfect for frying!

6. Groups

   Antecedents of plural pronouns can be non-contiguous. In that case, we follow the strategy explained below through example C.21.

   In example C.21, *your husband* and *Mrs. Humphries* constitute the antecedent for the plural pronoun *we*. But as they are non-adjacent markables, we can’t annotate them as a group. Therefore, we annotate them as markables but do not corefer with the pronoun *we*.

   (C.21) Did [[your]$_{i}$ husband]$_{j}$ buy Lorna, [Mrs. Humphries]$_{i}$? - No, [we]$_{k}$ bought her together.

7. Coordinated NPs

   Coordinated NPs are annotated both separately and as a whole

   (C.22) [[lies] and [assumptions]]

8. Numbers are annotated only if they are nominalized.

   (C.23) There were 100 participants in the meeting. [5]$_{i}$ among them was selected for the next step. [It]$_{i}$ was [an ambitious group]$_{i}$.

(C.24) [The first]$_i$ is [a woman]$_i$. [She]$_i$ will join us later.

9. Temporal Expressions are annotated.

   (C.25) We are going to meet on [Thursday]$_i$ because [it]$_i$ is [Anna's birthday]$_i$.

10. Do not annotate non-referential NPs in idioms, or lexicalized phrases such as "for example", "A penny for your thoughts" etc..

11. Predicative forms

    In simple copula relations, the mentions corefer. When a copula relation is negated, the mentions should not corefer.

    (C.26) (a) [Oxford]$_i$ is [a university]$_i$. [It]$_i$ has a long history.
    (b) [John]$_i$ is not [a lawyer]$_i$, [he]$_i$ is [an architect]$_i$.

12. Non-nominal referents (e.g. clauses, propositions, verbs) are not annotated. But the pronouns and nominal expressions referring these non-nominal antecedents are annotated and "referent_type" for these expressions should be selected as "clausal".

    (C.27) There is a big growth in the economy in last year. [This] is very surprising in current conditions.

    The antecedent for *This* is the whole sentence "There is a big growth in economy in last years.". In our annotation scheme, we do not annotate this sentential antecedent but we annotate "This" and assign the referent type "clausal" to this markable.

13. *One* pronoun

    "One" is annotated as a pronoun in the cases similar to the following:

    (C.28) "We have [one] (ellipsis: calendar) up for grabs.

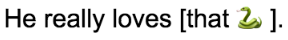
14. Usernames and hashtags

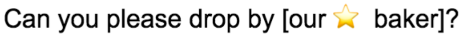
    In Twitter, there are automatically inserted usernames in the replies and also hashtags added by the users to increase the visibility of the tweets. The automatically inserted usernames (at the beginning of the tweet) and hashtags are not annotated unless they are referred by other expressions. The special cases in which the usernames and hashtags should be annotated are exemplified below:

    (C.29) [@BarackObama]$_i$ should change [his]$_i$ policy.
    (C.30) Yes! [She]$_i$ is my favorite. [#Oprah]$_i$

15. Emojis

    Emojis are annotated if they are used in place of nominals as in the examples below:

    (C.31) 
    He really loves [that 🐍].

    (C.32) 
    Can you please drop by [our ⭐ baker]?

We assign specific values to the *correct_form* and *comment* attributes in the scheme if the mention span contains an emoji:

- the field "comment" contains the string "emoji"
- The field "correct form" is for the string representation (e.g. "that snake" for the emoji-containing mention in C.31 and "our star baker" for C.32).

C.2.2 Spans of markables

Markables are always rooted in some nominal phrase (NP), and their extension is defined as follows:

- The syntactic head of the NP;
- Determiners and adjectives (if any) that modify the NP;
- Dependent prepositional phrases (for example, [Queen of England]).

Appositions, i.e., additive material that is not syntactically integrated are annotated separately (check example C.51 for this case). Nested mentions can also exist in the text, we allow the annotation of nested markable spans:

(C.33) In our colloquium today, [Slavoj Žižek]$_i$ will be talking about [[his]$_i$ new book]$_k$. [It]$_k$ is published by MIT Press.

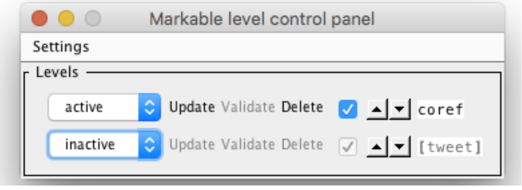
We do not allow discontinuous markables. Punctuations such as comma, paranthesis, question mark etc. are not involved in the markable span if they are not part of the proper name:

(C.34) He started to work in [Yahoo!].

C.3 Annotation Process

The annotation process selects all nominal expressions ('markables'). We annotate the complete reference chains for each entity referred by a nominal expression. Therefore, the annotation process involves a certain amount of "going back and forth" in the text.

If there are other annotation levels active in the scheme, the levels other than the "coref" level should be set to "inactive" to make the annotation process simpler as in the figure below:



Markable level control panel

Settings

Levels

active | Update Validate Delete ☑ ▲▼ coref

inactive | Update Validate Delete ✓ ▲▼ [tweet]

Annotation of coreference chains is an incremental process. First step is to highlight the referential nominal expressions (i.e., names, noun phrases and pronouns) in the text. This is done by selecting the "Create Markable on level 'coref"' option in MMAX. After all the mentions are highlighted in "coref" level, the ones referring to the same entity should be put in the same coreference chain. Each mention should be linked to the closest antecedent by selecting the "Mark as coreferent" option in MMAX. In case of cataphoric pronouns ('Before she left, Sue locked the door') the relation is to be established in forward direction (here: from 'she' to 'Sue'). The exophoric pronouns which refers to an entity out of the text (i.e., there is no overt antecedent in the linguistic domain) should also be annotated. In some cases, these exophoric pronouns may create singleton chains if there is no other mention in the text referring to the same entity as the exophoric pronoun. Non-referential (pleonastic) pronouns should also be annotated as markables and they will be considered as singleton coreference chains as no other mention is linked to them.

C.4 Attributes

C.4.1 Attributes for all markables

C.4.1.1 representative_men

With this attribute, we identify the most informative/descriptive mention as the representative mention of the entity. In general, here is the hierarchy in terms of representativeness among the mentions: NE>defNP>indefNP>pronoun. If there are more than one mention which can be considered as the representative of the coreference chain, select the first instance in the text. Every coreference chain should have only 1 representative mention.

C.4.1.2 np_form

1. none - not a nominal entity (no markable should be assigned to this type for this version of the guideline)
2. ne - named entity
3. defnp - definite NP
4. indefnp - indefinite NP
5. ppers - personal pronoun
6. ppos - possessive pronoun
7. padv - pronominal adverb
8. pds - demonstrative pronoun
9. prel - relative pronoun
10. prefl - reflexive pronoun
11. other - none of these options

Possessive pronouns mine/yours/his/hers/ours/theirs are coreferent with the possessed item, not the possessor (except for some special cases like in "friend of mine" where *mine* is coreferential with the other first person pronouns referring to the same speaker).

C.4.1.3 genericity

This value shows whether the referential expression under concern is referring to a specific entity or whether it is a generic nominal expression. Genericity is only assigned to the representative mention.

Please note that the plurals without a determiner, singular nouns with indefinite determiner (a/an) and NPs with "any/no" such as "no facts/no man/any person" are likely to be generic NPs. For all the NPs, it requires individual investigation of the NPs. There are also generic usage of pronouns as exemplified with the second person pronoun *you* in the examples below. Please note that the pronoun type for the generic pronouns is "exophora -> symbolic_deixis".

Nominal expressions are *generic* in the following cases:

(C.35) "[Parents]i_generic should take care of [their]i children."

(C.36) If [a man]i_generic says something like that,...

(C.37) Hmm, [you]i_generic can't really tell what has happened there. That incident is too complicated.

Singleton relative pronouns like this are annotated with genericity value "specific":

(C.38) you don't know [who] he's caring for.

C.4.1.4 grammatical_role

This attribute describes the grammatical role of the annotated mention or the grammatical role of the higher level nominal phrase where the annotated mention belongs to.

1. none - The mention is not part of the syntax.

   (C.39) We need to find [her]$_i$. [#ClaudiaJohnson]$_i$ (grammatical role for #ClaudiaJohnson should be chosen as *none*)

   (C.40) Yes! [She]$_i$ is my favorite. [@Oprah]$_i$ (grammatical role for @Oprah should be chosen as *none*)

2. sbj - The mention or the NP that this mention belongs to is a subject.
3. dir_obj - The mention or the NP that this mention belongs to is a direct object.
4. indir_obj - The mention or the NP that this mention belong to is an indirect object.
5. prep_phrase - The mention or the NP that this mention belongs to is a prepositional phrase.
6. copula_rel - The mention or the NP that this mention belongs to is part of the copula relation.
7. adv - The mention or the NP that this mention belongs to is part of an adverb.
8. other - none of these options

(C.41) Find [her]$_{\text{i\_directObject}}$! [#ClaudiaJohnson]$_{\text{i\_noGrammaticalRole}}$

(C.42) I gave [his]$_{\text{i\_directObject}}$ wallet to [her]$_{\text{i\_indirectObject}}$.

(C.43) Just check [[his] stats]: "his stats" is the "dir_obj", and the possessive pronoun before the noun is also marked as the same grammatical role ("dir_obj")

(C.44) Coming back to [his]$_{\text{i\_prepositionalPhrase}}$ house soon.

(C.45) [This]$_{\text{i\_subject}}$ is [[his]$_{\text{j\_copulaRelation}}$ favorite]$_{\text{i\_copulaRelation}}$.

(C.46) He is coming [today]$_{\text{i\_adv}}$.

The adverbs (if they are referring mentions) *here/there* are annotated as np_form–>padv and pronoun type will be "exophora -> symbolic deixis" if they are used to refer to the locations. For other adverbs like *home/today/yesterday...* np_form is assigned to defnp even some of them are deictic (e.g. the temporal adverbs today, tomorrow, next year etc.).

The grammatical role of reflexives could be not so clear as in the case below: Reflexive pronouns used for emphasis are annotated as appositives.

(C.47) [I]$_{\text{i\_subj}}$ did it [myself]$_{\text{i\_appositive}}$.

So for the reflexive pronoun, if it is clear that the pronoun is the object of the sentence, or prepositional phrase, assign the relevant grammatical role to the markable. But in all the other cases (as in the example above) assign the grammatical role "other" to the markables.

(C.48) [He]$_{\text{i\_sbj}}$ presents [himself]$_{\text{i\_dirobj}}$ as a change-maker.

(C.49) [He]$_{\text{i\_sbj}}$ cooked [himself]$_{\text{i\_indirobj}}$ a delicious cake.

(C.50) [He]$_{\text{i\_sbj}}$ prepared the food for [himself]$_{\text{i\_prepphrase}}$.

other_grammatical_role

More descriptive information on the grammatical type if the grammatical_role is selected as "other".

1. appositive - The referring expression is an appositive construction that modifies an immediately-adjacent noun phrase (which may be separated by a comma, colon, dash, or parenthesis).

2. vocative - The referring expression is a direct address to one of the participants in the conversation.

3. other - none of these options

(C.51) I called [Till]$_{\text{i}}$, [my friend]$_{\text{i\_appositive}}$, to invite [him]$_{\text{i}}$ to join us.

(C.52) [United Kingdom]$_{\text{i}}$ ([UK]$_{\text{i\_appositive}}$) has the world's fifth-largest economy. [It]$_{\text{i}}$ has a high-income economy.

(C.53) In case of situations such as "[you] [guys]", "[you] [fucking morons]": The pronoun is annotated as usual, "guys"/"fucking morons" as **defnp** and grammatical role as "appositive" and they are marked as coreferent.

(C.54) Hey [@lynda]$_{\text{i\_vocative}}$, are you going to join us today?

C.4.1.5 semantic_class

For the sake of simplicity, this attribute is assigned only to representative mention in a coreference chain.

1. none - The mention is a non-referential pronoun.

2. abstract - The mention refers to an abstract concept.

3. human - The mention refers to a human, including fictional characters.

4. org - The mention refers to an organization.

5. loc - The mention refers to a location.

6. pyhs_obj - The mention refers to a physical object.

7. event - The mention refers to an event. (e.g. *hurricane, heart attack* etc.)

8. time - The mention refers to a certain time.

9. other - none of these options

(C.55) [True love]$_{\text{i\_abstract}}$ is rare, [it]$_{\text{i}}$'s [the only thing that gives life real meaning]$_{\text{i}}$.

C.4.1.6 pronoun_type

1. none - The interpretation of the expression does not depend on other mentions in or out of the text.

2. non-referential pronoun - These pronouns are semantically empty, and so, refers to no entity.

   For instance, *it* pronouns in the following examples should be marked as non-referential pronouns.

   (C.56) It's raining. (weather)

   (C.57) It takes 4 hours to go to Minneapolis. (time)

   (C.58) It seems that John is a good football player. (usage with a raising verb "seem", e.g. appear, look, mean, happen)

   (C.59) It is known that... (usage with a cognitive verb, e.g. think, believe, know, anticipate, recommend etc.)

   (C.60) It is clear that we should decline...

   (C.61) You can make it! (part of the idiom, make it=succeed, e.g. on the face of it)

3. anaphora - The expression refers to a backward phrase in the text.

4. cataphora - The expression refers to a forward phrase in the text.

5. exophora - The expression refers to extra linguistic context.

6. bridging - A definite NP picks up some aspects of a previously introduced referent and enters into a relation with that referent other than identity. (This attribute is not used in the scope of this version of the guideline but we kept it for future changes in the annotation scope.).

exophora_type

Type of the exophoric mention

1. symbolic_deixis - Pronouns point to a referent not inside the text but in the situation of utterance (e.g. spatio-temporal or speaker knowledge is required to interpret the pronoun). Usually first and second person pronouns are considered as symbolic deixis. We annotate the first occurence of these pronouns **in every chain** as exophoric deictic but the other pronouns referring to the same entity, and so belonging the same coreference chain, are marked as anaphoric.

2. antecedent_in_attached_picture - The antecedent of the pronoun is not in the linguistic context but in the visual media attached to the text.

3. antecedent_in_attached_text - The antecedent of the pronoun is not in the current linguistic context but in the text pointed by the link attached to the text.

4. antecedent_in_quoted_tweet - The antecedent of the pronoun is not in the linguistic context but in the quoted (embedded) tweet.

5. antecedent_inferred_by_world_knowledge - The antecedent of the pronoun is not in the linguistic context but can be inferred by world knowledge.

referent_type

Type of the referent (referred expression)

1. nominal - The referent of the pronoun is a nominal entity.
2. clausal - The referent of the pronoun is a clausal entity. As we only annotate nominal referential expressions, the clausal referring expressions are not annotated. Only the nominals referring to these clauses can be annotated.

   (C.62) John didn't call me yesterday. [This]$_{\text{i}}$ made me sad.

   In the example above, the pronoun *This* refers to the whole clause *John didn't call me yesterday.* But we don't annotate the clausal expressions. Therefore, *This* is annotated but its clausal antecedent is not annotated.
3. other - none of the above

C.4.1.7 correct_form

If there is a typo or misspelling in the surface form of the mention, the correct spelling is written here by the annotator.

C.4.1.8 comment

We use this attribute to add more information about the mentions. This field is available for free comments but we also use is to add some more features about the annotated mentions, if necessary. Instead of adding new attributes for the features below, we use this field to add more information on the markable:

- **metadata**: If the mention is not part of the syntax but instead part of the conversation or message structure offered by the communication media such as usernames automatically added to the replies. Please note that the "grammatical_role" should be selected as "none" if the "comment" is set to "metadata" value.

  (C.63) *(Tweet_1)* **@StarTimesKenya:** [@dennisclaude89]$_{\text{i\_metadata}}$ [you]$_{\text{i}}$ should downgrade your account.
  (Reply_to_Tweet_1) **@splinister:** @StarTimesKenya @dennisclaude89 I don't think [he]$_{\text{i}}$ needs to downgrade...

  Please note that in the example above, only the first instance of **@dennisclaude89** is annotated, the other instances which automatically added to all the replies will not be considered as markables.

- **emoji**: If the annotated mention span contains an emoji sign, the string "emoji" is assigned to comment field.

C.5 Interesting cases

- In the constructions similar to following examples, the name is marked as coreferent with the person/pronoun but not with the NP "name": [[My]$_{\text{i}}$ name]$_{\text{j}}$ is [Jordan Smith]$_{\text{i}}$.
- "bro", "dude", "dear" etc. are annotated as npform "defnp" and grammatical role "other > vocative"

- NP expressions like “such usage”, although they are anaphoric, are annotated as NP and without a pronoun type.

- “I am skeptical of [adjectives being used without [their nouns]]”: both are PPs, recursive/nested grammatical roles are annotated as what they are (regardless of the recursion).

- Speaker A: What [a liar]$_i$
  Speaker B: “R u calling [me]$_j$ [a liar]$_i$?
  Both “a liar” are marked as coreferent, but not coreferent with “me” (as it is contentious)

- “[Florida representative]$_i$[Matt Gaetz]$_i$”: The name “Matt Gaetz” is the representative mention, “Florida representative” is marked as “other > appositive”

- “[Indian Hindu] support [Isreal]”: Both are marked as “org” because they refer to political groups/governments.

- “Let [’s] talk about”: “’s” (us) is annotated as “generic” and “ppers”.

- “you don’t know [what] [their intentions] are”: “what” is annotated as “prel” and it is a singleton.

- “if [he]$_i$ said outright ’[I]$_i$ don’t like gay marriage”’: The pronouns are coreferent, regardless of “if” and direct speech.

- “In [2018]”, “In [March]” ... : Years and months are annoated as named entities (i.e., np_form *ne*).

- The relative pronoun “what” can take on many pronoun_types:
  - it may be exophora -> antecedent_inferred_by_worldknowledge if the referent is known and outside the text (“[what] Obama said”),
  - anaphoric if it refers to a already mentioned referent (nominal: “[This]i is [what]i I mean”) or the antecedent can be derived from context (clausal: “[What] you said here”)
  - cataphoric in sentences like “[What]i I need is [time]i” (nominal) or “[What] they talked about was how ...” (clausal)
  - If “what” introduces a new antecedent, pronoun_type might be *none* for the first mention (“I saw [what] you cooked”)

- “both” is annotated as np_form “other”.

- Here/There: Every occurence is annotated for consistency and as “symbolic_deixis” in the first and/or representative mention. If there is another here/there in the same chain, it is annotated as “anaphora”. To differentiate between local here/there (“here in Australia”) and other uses (“Yet here we are” (in this situation), “Here on Twitter”), the first is annotated as semantic_class “loc”, the non-local use as “other”.

C.6 Validation Checks

The data is automatically validated in the following ways, with the help of scripts:

- All the chains (including singletons) have 1 representative mention.

- semantic_class and genericity are assigned to all representative mentions (and not to any other mention)?
- only the first occurrence of deictic expressions are marked as exophoric, all the other mentions referring to the same entity are marked as anaphoric.

C.7 Version Tracking

We keep track of the revision details of this guidelines in this section.

Version	Description
v1.0	The first version of the guidelines. This version describes the instructions for annotating the complete reference chains containing at least one third person singular pronoun.
v2.0	This version describes the instructions for annotating all referential expressions and reference chains in the text.
v2.1	The decisions made during the annotation process of complete chains are added to the guideline.

Table C.1: Version details

Appendix D

Coherence Relations Annotation Guideline for TwiConv

D.1 Introduction

This guideline provides the basics for the annotation of explicit, implicit and hypophora relations in the genre of Twitter conversations. Twitter[1] is a social media platform that publishes short "microposts" by registered users. In addition to textual content, these posts may contain embedded images or videos. Twitter users can interact by directly replying to each other's messages. Such replies are quite frequent and the resulting conversations often contain discourse connectives [Scheffler, 2014].

D.2 General Annotation Procedure

We follow the general PDTB framework for annotating the Twitter conversation threads, with some adjustments that are necessary to deal with the specifics of Twitter conversations.

A good starting point would be to have a look at the PDTB 3.0 annotation manual [Webber et al., 2018], particularly the explicit, implicit and hypophora senses and the list of connectives in the last section, in order to have an idea of what expressions can potentially (but not necessarily) qualify to be connectives in our data.

The argument spans for all relations are not previously decided or marked, so additionally to the relation itself, the annotator has to choose which span boundaries are meaningful. This can be complicated by the non-standard language of Twitter, meaning that the requirements for a valid span are sometimes different than in the original PDTB annotation manual Webber et al. [2018] (c.f. D.3.1 for span selection details).

We annotate all explicit connectives; in case of implicit relations the connective is chosen by the annotator (with the help of a list of possible connectives). For each relation, the two arguments are identified and the connective sense is labelled according to the PDTB 3.0 relational taxonomy[2]. We primarily used the list of 100 explicit connectives from the PDTB corpus [Prasad et al., 2008] to identify connectives. Additionally, we found a few new connectives in our corpus such as *by the way, plus, so long as*, and *when-then*. If we find an ambiguous connective or interpret more than one relational reading, we assign multiple senses to the connective. For hypophora relations, only the argument spans are labelled (Arg1 for the question, Arg2 for the response). The annotation was conducted

[1] `www.twitter.com`

[2] We do not annotate other information such as attribution features or supplementary spans for connectives.

using the PDTB annotator tool (PDTB Annotator v4.6[3] and v4.9).

D.3 Annotation Decisions

In this guidelines, Arg1 is marked by *italic letters*, Arg2 by **bold letters** and connectives by underlining. Some key points that we jointly discussed and agreed on are as follows:

- Relations and their connectives can occur in intra-tweet (within a single tweet) or inter-tweet contexts (across tweets). Some authors may compose their messages in more than one tweet, i.e., they begin with one tweet and continue writing in the subsequent (adjacent) tweets, or a user may connect his own message with a previous (adjacent) message by another author.

  (D.1) *Black folks in Alabama organized.* And **WON!** [Single Tweet]

  (D.2) *Like I said, you don't know the whole situation to make such a judgement.* [Tweet1]
  And **until you have raised one yourself, sit down and shut up!** [Tweet2]

- Only *Arg1*, *Arg2* and the explicit or implicit connectives are tagged (and nothing else, such as attribution features/source or supplementary spans).

- Up to two senses can be annotated if necessary for implicit and explicit relations.

D.3.1 Argument Spans

When deciding on a suitable relation and sense, it is important to pay attention to the argument span boundaries. Not only can including more or less material in a span affect the sense, but which constructions qualify as spans differs from the original PDTB guidelines as well. According to them, arguments are usually clauses, in addition to some other construction types such as nominalization and VP conjuncts (see the PDTB annotation manual). We have applied more flexible parameters for determining arguments.

1. Generally, as in the original PDTB we aim to catch the sense that is most likely intended to be read by the author. If the meaning of a clause or sentence can be ambiguous, the chosen span should reflect which meaning or sense the annotator thinks is most probable. It can be helpful to choose the shortest possible span that still makes sense first, and then decide whether adding more context changes the meaning.

   (D.3) According to him *everything was perfect* but **it was raining the whole day.**
   In this case, ...

   (D.4) *According to him everything was perfect* but **it was raining the whole day.**
   By including *According to him* ...

   It is also possible to have a relation with two arguments, and those two arguments together make a new Arg1 or Arg2 for another relation.

[3] https://drive.google.com/file/d/1b3n7CDLoT1bPxkp5lHLC_kEHVCUNsaEk/view

(D.5) ...

2. Tweets often represent fragments of utterance, and may not necessarily be represented as clauses. In addition to clauses and other standard argument types, elliptical constructions (with an implied verbal predicate) may be identified as arguments. Therefore PPs, VPs or NPs can be arguments.

(D.6) *Dm me* if **so**
Ellipsis test: ... if **(it is) so**

(D.7) *only second year* and **got much more upside than Bayliss**
Ellipsis test: *(It is) only (his) second year* and ..

(D.8) *WE ARE NOW* AND **THANK YOU**
Ellipsis test: ... AND **(I) THANK YOU**

(D.9) *Do you know if the loverly Gary Barlow is doing a keepfit dvd* **thanks** anyway

(D.10) *Typical misogynistic behaviour on ur part* but **not entirely unexpected**

(D.11) *NO PROB* BUT **WHERE THE HELL DID U**

(D.12) If **he could work on that,** *good prospect.*

3. Twitter data contains a wide range of acronyms for fixed expressions, such as 'idc' = 'I don't care'; 'idk' = 'I don't know'; 'idrk' = 'I don't really know'; 'tbh' = 'to be honest'. The annotator needs to decide on whether these acronyms function like arguments or not in particular contexts.

(D.13) *idc* if **u do or not**

(D.14) *why* when **ordinary folk are being made hungry and homeless tbh**

4. Arguments can be discontinuous if they contain unnecessary/unrelated material (e.g. clauses in parentheses, other clausal structures, hashtags) or the structure requires it.

(D.15) Speaker 1: *Yes or no, is it unreasonable to expect immigrants to conform to our laws and not form crime gangs?* (Obviously this doesn't concern all immigrants or any from a particular race, just a basic question) → Material in parentheses is omitted from Arg1
Speaker 2: **Mostly. I'm giving an answer that you will no doubt see as a cop out because not everything is a binary.**

(D.16) *which* by the way *is not against the law* and **there is still no evidence**

(D.17) *And when our immigration system is broken and Mexicans are entering the country illegally, some "bad hombres" are slipping in too.* #commonsense And **who cares if the guy has a sombrero on.** → Hashtag is omitted

However, to keep spans coherent and and the annotation process simple, short material, e.g. single adverbs, common phrases, vocatives, interjections etc. do not necessarily need to be excluded.

(D.18) *why* when **ordinary folk are being made hungry and homeless tbh** → 'tbh' is not omitted

(D.19) *like i wanted to defend CWC but then i saw they put on blackface to be funny* and **uh a bit harder now** → 'uh' is not omitted

5. Rarely, in hypophora relations, URLs can be considered argument spans (Arg2) if the content the URLs directs to is meant as a response to the question asked.

   (D.20) Speaker 1: *What has Blair Cottrell said that is racist?*
   Speaker 2: **https://t.co/NcUANEPttZ**

6. Sometimes authors deliberately leave out arguments although they use an explicit connective, therefore making the relation incomplete. The PDTB annotator tool will mark the annotation red, but we decided to allow such incomplete relations.

   (D.21) If only *Tory elitism were* (omitted: **then ...**)

   (D.22) *He 's gone now* so (omitted: **we will never know**)

7. Arguments can span across tweets. The arguments are then treated as discontinuous spans, non-argument material between them is omitted (i.e., the author's and/or other automatically inserted username)

   (D.23) *It's less about expecting a woman you give emotional support to to become a romantic partner* but **rather the advice to not become a friend. . .** [Tweet1]
   . . . of a woman you're romantically interested in [Tweet2]

   (D.24) *she had the right to call him out* because **he was+** [Tweet1]
   +careless with his words, no matter who he is. [Tweet2]

8. If a span has some final punctuation (which is not always the case in Twitter messages), the punctuation it is part of the argument span. If a punctuation character is repeated several times, all repetitions are included in the span.

   (D.25) *I love that look of satisfaction on her face* when **she figures something out!!!**

9. Emojis added at the end of an argument are left out. If the emojis are syntactically integrated or otherwise essential to the understanding, they should be part of the span.

   (D.26) *We need animated emojis* bc **the [running man emoji] is great but imagine if my guy was actually grooving**

   (D.27) *First of all, I 'm a virgin, thank you very much.* (Implicit=And)
   Secondly, I don't know what you're taking about [laughing face emoji]

D.3.2 Relation Types

Three relation types are annotated - Explicit, Implicit and Hypophora. AltLex, AltLexC, EntRel and NoRel are never annotated.

D.3.2.1 Explicits

Explicit relations can be coordinating or subordinating conjunctions between two spans (as they are defined in section D.3.1). PDTB 3.0 also considers prepositions such as 'because of', 'with' or 'despite' as connectives. While we allowed more flexible spans than the strict clausal definition in the PDTB guidelines, it is still important to check whether such prepositions do not only connect to a pure nominal phrases (e.g. 'It was cold despite the sunshine' is only one span). It is helpful to confirm that each span could be read as a clause on its own, which often means checking if it contains a verb (or if is elided and could be inserted).

- Twitter data often shows variant forms/spellings of a connectives such as 'wen' = 'when'; 'cos', 'cus', 'cuz' = 'because'; 'btw' = 'by the way'; '&', '&amp', 'an'= 'and'. These variants will be annotated if relevant.

  (D.28) Omve **he starts hitting them** *the game plan goes out the window* → 'Omve' (once) is a connective

- Twitter may also contain unusual connectives not included in the PDTB corpus.

  (D.29) *Kohli scored a lot of runs in Australia right?* Plus **this Test was hardly a blowout** → 'plus' is a connective.

  (D.30) *Oprah is* & **can do that.** → '&' is a connective.

- Double connectives like 'but then', 'but also', 'and then' are always annotated separately, even if they have the same sense. If it is impossible to assign separate senses to one or both components, a comment should be added.

  (D.31) *i really liked them together* but also **i wouldnt mind seeing emma and robert dating.** (but=Comparison.Concession.Arg2-as-denier, also=Expansion.Conjunction)

D.3.2.2 Implicits

The general annotation strategy is to first try inserting (different) possible connectives between spans, before settling on one and then annotating the sense (or two, if required). Often, multiple connectives are possible. In this case, the annotator has to decide which connective reflects the most natural (and most likely author-intended) relation and make sure that in case of two possible senses, the connective reflects both of these senses. As an example, 'and' can very often be inserted between spans while sounding natural, but it does not clearly show any causal or other intended relations. It can be helpful to use the list of connectives in the appendix of the PDTB manual as inspiration for possible insertions.

Even with careful thought and trying different options, the exact implicit connectives can differ for each annotator, but be semantically equal and be annotated with the same sense (e.g. 'since' and 'because' can often be used interchangeably for causal relations).

- When a clause is followed by a listing of examples or instances referring to that one initial clause, instead of linking each clause to the initial one, the list is "chained" and only neighbouring clauses are annotated.

  (D.32) **UK saw a surge in GDP**. (Implicit=) *Manufacturers report fullest order books for 30 years.*
  Manufacturers report fullest order books for 30 years. (Implicit=) *Strategic investments from Boeing, Hitachi, Tata, Liberty, Honda & Toyota.*
  Strategic investments from Boeing, Hitachi, Tata, Liberty, Honda & Toyota. (Implicit=) *FTSE100 closed at record level 29/12/17 & so did FTSE250.*

- Sometimes adding an implicit connective can cause an argument to sound unnatural because of the Twitter-specific language (e.g. dropping the subject). In case of simple ellipsis, the relation should still be inferred.

  (D.33) *Not surprised watching India get beaten in SA* (Implicit=as) **[I] never thought much about them away from home.** ["I" is dropped by the author]

D.3.2.3 Hypophora

Hypophora is defined as question-response pairs, where Arg1 is the question and Arg2 the response. A 'response' in this case is anything that aims to answer or otherwise contributes to the first argument, even if not directly providing the requested information (e.g. pointing out that a question cannot be answered due to lack of knowledge or the question being irrelevant, counter questions etc.). Rhetorical questions (which by default should not have a response/Arg2) are not annotated with the Hypophora relation.

- The two arguments (question and response) can be from either the same speaker or two different speakers.

  (D.34) Speaker 1: *What was the film?*
  Speaker 2: **Jupiter's moon**

  (D.35) *Why don't you have outrage for this?* Oh I forgot. (Implicit=Because) **It's another failed attempt to deflect.** [Same speaker]

  (D.36) You probably voted to leave to stop immigration and *you know what?* **It won't.** [Same speaker]

- If there is an explicit connective or an implicit relation can be inferred, the same argument spans are annotated once as Explicit/Implicit and once as Hypophora.

  (D.37) Speaker 1: *why you angryyy?*
  Speaker 2: (Implicit=Because) **I just don't like teams stacking up like that**

D.3.3 Comments

Comments can be added for irregular occurrences, e.g.:

- Sometimes it is not possible to select the correct span or connective because of incorrect tokenization. The correct spans should be mentioned in the comment.
- If paired connectives cannot be separated and be assigned separate senses, it should be pointed out.

D.3.4 Interesting Cases and Specific Decisions

- Concession and contrast relations can sometimes be signalled by the same connectives and therefore be confused in both explicit and implicit uses. Contrast should only be annotated when the arguments spans have similar structures (syntactically and/or semantically) which contrast with each other, while Concession "simply" means that one argument denies the other. The PDTB3 annotation manual Webber et al. [2018] recommends the following strategies to distinguish them:

  1. Are at least two explicit differences highlighted between the arguments?
  2. If no, select Concession. 23
  3. If yes, check whether a causal relation that is expected on the basis of one argument is denied by the other (Test by paraphrasing with 'although').
  4. If yes, select Concession.
  5. If no, select, Contrast.

  (D.38) *Max has a red car,* (Implicit=but/while/... Comparison.Contrast) **Susan has a white one.**

(D.39) *Max has a red car*
but/although/... (COMPARISON.CONCESSION.ARG2-AS-DENIER) **he doesn't like the colour at all.**

- *that's because* and similar causal connectives such as *that's coz* are treated like connectives, but the similar appearing "That's when" is not (which would an AltLex phrase, but those are not annotated here).
- The abbreviation *i.e.* (id est, that is to say) can be a connective (e.g. Expansion.Equivalence) if it connects proper argument spans, although there was no instance found in this corpus so far.
- *anyway* might also be a connective, and can be checked via substitution tests (*regardless*, *in any case*)
- *So, ...* as well as some other markers, especially at the beginning of the conversation or tweet, can be a connective but might also be a topic starter/changer and not necessarily indicate a coherence relation.

D.4 Corpus and Tool Versions

The explicit and implicit relations have been annotated with the PDTB Annotator v4.6 until January 2020. All newer annotation updates as well as hypophora relations are annotated with v4.9.

D.5 List of Connectives

Connectives annotated in this study:

and

but

if

so

when

because

also

&

or

as

though

then

before

while

until

however

even if

without

since

like

if .. then

yet

after

unless

instead of

still

cos

by

because of

bc

anyway

tho

in case

whilst

so that

rather

once

not only .. but

meanwhile

instead

even when

even though

cuz

btw

as well

an

whereas

whenever

wen

rather than

plus

otherwise

nor

just because

furthermore

for

except

either .. or

but then

b4

as soon as

as long as

and then

although

where as

where

when .. then

when .. if

whatever

till

thus

thou

therefore

thereby

that when

that is why

so long as

so as to

so as

regardless of

regardless

omve

like when

just becasue

in case

in the meantime

if .. than

if only

if not .. then

if not

hence

given

for when

for example

every time

even after

despite

cus

coz

cause

by the way

besides

becasue

as a result of

anytime

and therefore

and btw

ah

after all

+

Appendix E

CONLL-formatted TwiConv Data

```
1_987654321000000000.branch1.); part 0
000.branch1.  0  0  This  DT  (ROOT(S(NP*)  SomeUsername  -  -  pds/cataphora  (0)  CL0  NP_S(  NP_L(  -  -  0_0_0
000.branch1.  0  1  is  VBZ  (VP*  SomeUsername  -  -  -  -  -  -  )NP_S  -  -  -  0_0_1
000.branch1.  0  2  just  RB  (ADVP*  SomeUsername  -  -  -  -  -  -  -  -  -  -  0_0_2
000.branch1.  0  3  a  DT  (NP*  SomeUsername  representative_men  -  indefnp/none  (1  -  NP_S()  )NP_L  -  -  0_0_3
000.branch1.  0  4  test  NN  *)))  SomeUsername  -  -  -  1)  -  -  -  -  -  0_0_4
000.branch1.  0  5  .  .  *))  SomeUsername  -  -  -  -  CL0  -  -  -  -  0_0_5

000.branch1.  0  0  Hi  UH  (ROOT(INTJ*  SomeUsername  -  -  -  -  CL1  -  -  -  -  0_1_0
000.branch1.  0  1  Twitter  NNP  (NP*)  SomeUsername  representative_men  -  ne/none  (2)  -  -  -  -  -  0_1_1
000.branch1.  0  2  !  .  *))  SomeUsername  -  -  -  -  CL1  -  -  -  -  0_1_2
```

Figure E.1: A Sample TwiConv CoNLL file

Appendix F

Stimuli in Story Continuation Experiment

- 1-clause length

    1. John collected samples for investigation.
    2. Emily created an Instagram profile.
    3. Michael worked hard for weeks.
    4. Patricia rode all day yesterday.
    5. Robert finally solved the mystery.
    6. Linda fell in love again.
    7. Mary received good news today!
    8. James finally found the keys.

- 2-clause length

    1. William painted a picture. It will be a perfect present.
    2. Jennifer remembered that night when the same smell came from outside.
    3. Richard stayed at home because it had been raining heavily outside.
    4. Thomas recognized everything immediately. Nothing had changed during these years.
    5. Margaret jumped into the water before the old boat sank completely.
    6. Susan founded a new startup company, then unfortunately the pandemic started.
    7. Barbara had a quick lunch before the doorbell rang once again.
    8. David packed everything into a backpack. It is finally over!

- 3-clause length

    1. Cristopher attended a meeting in London. It was about climate change. The discussions were fruitful.
    2. Lisa washed the dishes resentfully after the party ended. It was a long and gloomy night.
    3. Paul baked a cake today. The house feels like home now because the cake smells amazing.
    4. Mark smiled with relief and pride. The launch was successful. The project is completed now.

5. Betty slept early. It was a quiet night until the phone exploded violently under the bed.
6. Nancy became famous after the movie was released online. It garnered great attention on social media.
7. Dorothy prepared stimuli for the experiment. This experiment will be expensive because it lasts too long.
8. Daniel grabbed a coffee. There was too much work. Fortunately, Mondays are easier with caffeine.

- 5-clause length

  1. Matthew went out for the first time after the lockdown was lifted. Restaurants were open and busy. The streets were empty because it was very cold outside.
  2. George posted a tweet today. Social media is interesting. It can be both dark and bright. It was beneficial this time because the message reached millions.
  3. Laura nervously opened the letters. One was from the bank. Another one came from the school. The others were advertisements. This last one seems important.
  4. Doug was disappointed when the launch of the rocket was cancelled. Everything went well until a strong wind suddenly came up. The weather was so unpredictable here.
  5. Michelle prepared the food cheerfully. The first dish was salad. The second was a Green Curry. They looked very tasty. Everything was ready for dinner.
  6. Carol made a peppermint tea before the sun rose. Herbs can balance the emotions. Peppermint helps for anxiety so it was a perfect choice for this morning.
  7. Catherine first played the song at a festival. It was a huge event. The venue was full because admission was free. The song became famous afterwards.
  8. Jack booked a train to Vienna. The station is crowded because the holiday season has started. The train departs at noon, so there is time for coffee.

- 8-clause length

  1. Danny moved to an island. It was rather small. Vehicles were not permitted there. It was a mysterious place, too. There was a volcano in the South. One night, the volcano became active, then the sky brightened. That was really scary.
  2. Nicolas became vegan. It was a political decision because animal agriculture affects the climate. It was also a health matter. Animal products can be harmful because antibiotics are used in farming. Veganism also supports weight loss. This was actually the main reason.
  3. Theresa published an article in Lancet. Lancet is a leading medical magazine. It covers everything about health. It publishes up-to-date studies. There is a review process before the articles are issued. Lancet accepts only original papers. Therefore, this is an impressive success.
  4. Shannon fell asleep on the couch while the TV was on. CNN was reporting the news. It was about the war. The war videos were clear because satellite technology had improved significantly. The sound was also clear. It was breaking the silence.

5. Frederic took the train to Paris. It departed from Waterloo station. The trip took 3 hours. The weather was amazing there. Autumn was starting so the trees were almost yellow. The city was really beautiful. It also seemed generally very peaceful.
6. Charlotte was tired. There are good days in life; there are also challenging days. Today was one of the latter. There was nobody around. The building was almost empty. It was getting dark. Even the computer turned to sleep mode.
7. Gabriela moved to a new neighborhood. The previous area was chaotic because it was very central. This new neighborhood is quiet. There are parks nearby. The new house is great. It is quite peaceful. A whole new life can start here.
8. Tony started organic farming for health reasons. Organic food is healthy because no chemicals are used in the process. It yields more food with less expense when it is done systematically. The products are much tastier: Tomatoes smell wonderful, peppers are flavorful.

- Filler (control) stories

1. Deborah noticed that she really missed her college days.
2. I woke up very energetic today and this is quite exceptional.
3. The hitcher seemed scruffy in his parka and Pink Floyd t-shirt.
4. “It’s almost three o’clock” Ms. Jones whispered accusingly.
5. Kevin was very lucky: He managed to make serious cash from the lottery.
6. Jerry is looking at the photos on Victoria’s iPad.
7. After walking two or three times along that part of the lane, Elizabeth was tempted, by the pleasantness of the morning, to stop at the gates and look into the park.
8. I wonder why all this was not told me last night.
9. The evening passed quietly, unmarked by anything extraordinary. At night, Darcy opened his heart to Jane.
10. Dill left us early in September, to return to Meridian.
11. The door opened quietly, the dark silhouette of a man entered the room, and the door closed again. The man seemed to be looking for something.
12. He pointed to the east. A gigantic moon was rising behind the trees.
13. Anthony was an ordinary man. One thing that he did not believe in was superstition.
14. Miss Caroline began the day by reading a story about cats.
15. I must now mention a circumstance which I would wish to forget myself.
16. Walter had picked himself up and was standing quietly listening to the music.
17. There once was a bored shepherd boy who longed for more excitement in life while dutifully watching the flock of sheep on the hill.
18. One evening, the woman went to the theatre with two friends and saw an interesting guy there.
19. Miss Smith hated her house: time spent indoors was time wasted.
20. I married early and was happy to have a family.

Bibliography

Syed Muhammad Faisal Abbas. *Microblog text parsing: A comparison of state-of-the-art parsers*. Dalhousie University, 2015. URL https://books.google.de/books?id=fND9twEACAAJ.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-0705.

F. Niyi Akinnaso. On the differences between spoken and written language. *Language and Speech*, 25:125 – 97, 1982.

Berfin Aktaş and Annalena Kohnert. TwiConv: A coreference-annotated corpus of Twitter conversations. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 47–54, Barcelona, Spain (online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.crac-1.6.

Berfin Aktaş, Tatjana Scheffler, and Manfred Stede. Anaphora resolution for Twitter conversations: An exploratory study. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–10, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0701. URL https://aclanthology.org/W18-0701.

Berfin Aktaş, Veronika Solopova, Annalena Kohnert, and Manfred Stede. Adapting coreference resolution to Twitter conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2454–2460, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.222. URL https://aclanthology.org/2020.findings-emnlp.222.

Berfin Aktaş and Manfred Stede. Variation in Coreference Strategies across Genres and Production Media. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

Marilisa Amoia, Kerstin Kunz, and Ekaterina Lapshinova-Koltunski. Coreference in spoken vs. written texts: a corpus-based analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the*

Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Anietie Andy, Chris Callison-Burch, and Derry Tanti Wijaya. Resolving pronouns in Twitter streams: Context can help! In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 133–138, Barcelona, Spain (online), December 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.crac-1.14`.

Chinatsu Aone and Scott William. Evaluating automated and manual acquisition of anaphora resolution strategies. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129, Cambridge, Massachusetts, USA, June 1995. Association for Computational Linguistics. doi: 10.3115/981658.981675. URL `https://aclanthology.org/P95-1017`.

Mira Ariel. *Accessing Noun-Phrase Antecedents.* Routledge, 1990.

Jennifer E. Arnold. *Reference form and discourse patterns.* PhD thesis, Stanford University, 1998.

Jennifer E. Arnold and Zenzi M. Griffin. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, 56(4): 521–536, 2007. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2006.09.007. URL `https://www.sciencedirect.com/science/article/pii/S0749596X06001380`.

Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL `https://www.aclweb.org/anthology/J08-4004`.

Nicholas Asher and Alex Lascarides. *Logics of Conversation.* Cambridge University Press, Cambridge, 2003.

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566, 1998.

Hongxiao Bai and Hai Zhao. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://aclanthology.org/C18-1048`.

David Bamman, Olivia Lewke, and Anya Mansoor. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May 2020. European Language Resources Association. URL `https://aclanthology.org/2020.lrec-1.6`.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

BBN Technologies. *Co-reference Guidelines for English OntoNotes Version 7.0*, 2007.

David Bean and Ellen Riloff. Unsupervised learning of contextual role knowledge for coreference resolution. In *Proceedings of the Human Language Technology Conference* *of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 297–304, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL `https://aclanthology.org/N04-1038`.

Shane Bergsma and Dekang Lin. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, page 33–40, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220180. URL `https://doi.org/10.3115/1220175.1220180`.

Douglas Biber. *Variation across Speech and Writing.* Cambridge University Press, 1988. doi: 10.1017/CBO9780511621024.

Douglas Biber. Using computer-based text corpora to analyze the referential strategies of spoken and written texts. In Jan Svartvik, editor, *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, pages 213–252. Berlin:Mouton, 1992.

Douglas Biber and Susan Conrad. *Register, Genre, and Style.* Cambridge Textbooks in Linguistics. Cambridge University Press, 2 edition, 2009. doi: 10.1017/9781108686136.

Douglas Biber, Edward Finegan, Stig Johansson, Susan Conrad, and Geoffrey Leech. *Longman Grammar of Spoken and Written English.* Longman, 1 edition, 1999.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc., 2009.

Daniel G. Bobrow. Natural language input for a computer problem solving system. Technical report, USA, 1964.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *In SIGMOD Conference*, pages 1247–1250, 2008.

Arnoud Boot, E. (Erik) Tjong Kim Sang, Katinka Dijkstra, and Rolf Zwaan. How character limit affects language usage in tweets. *Palgrave Communications*, 5(1), December 2019. doi: 10.1057/s41599-019-0280-3.

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, ACL '87, page 155–162, USA, 1987. Association for Computational Linguistics. doi: 10.3115/981175.981197. URL `https://doi.org/10.3115/981175.981197`.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419, 12 2010. ISSN 1574-020X. doi: 10.1007/s10579-010-9120-1.

Yang Trista Cao and III Daumé, Hal. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, 47(3):615–661, 11 2021. ISSN 0891-2017. doi: 10.1162/coli_a_00413. URL `https://doi.org/10.1162/coli_a_00413`.

Claire Cardie and Kiri Wagstaff. Noun phrase coreference as clustering. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999. URL `https://aclanthology.org/W99-0611`.

Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, June 1996. ISSN 0891-2017.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32550-5.

Wallace L. Chafe. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li, editor, *Subject and topic*, page 25–55. Academic Press, New York, 1976.

Wallace L. Chafe. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production.* Cambridge University Press, 1980.

Wallace L. Chafe. Integration and Involvement in Speaking, Writing and Oral Literature. In Deborah Tannen, editor, *Spoken and Written Language: Exploring Orality and Literacy*, pages 35–54. Ablex, Norwood, NJ, 1982.

Wallace L. Chafe. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* University of Chicago Press, 1994.

Wallace L. Chafe and Joanna Danielewicz. Properties of spoken and written language. In *Comprehending oral and written language*, pages 83–113. Academic Press, San Diego, 1987.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1016. URL `https://aclanthology.org/D18-1016`.

Christian Chiarcos and Olga Krasavina. Rhetorical distance revisited: A pilot study. In *Proceedings of the Corpus Linguistics Conference*, Birmingham, 2005.

Sofiana-Iulia Chiriacescu. *The discourse structuring potential of indefinite noun phrases. Special markers in Romanian, German and English.* PhD thesis, University of Stuttgart, 2011.

H. Clark. Bridging. In P. N. Johnson-Laird and C. Wason, editors, *Thinking: Readings in cognitive science*, page 411–420. Cambridge University Press, Cambridge, 1977.

H. H. Clark and C. J. Sengul. In search of referents for nouns and pronouns. *Memory and Cognition*, 7:35–41, 1979.

Kevin Clark and Christopher D. Manning. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1061. URL `https://aclanthology.org/P16-1061`.

Kevin Bretonnel Cohen, Arrick Lanfranchi, Miji Choi, Michael Bada, William A. Baumgartner, Natalya Panteleyeva, Karin M. Verspoor, Martha Palmer, and Lawrence E. Hunter. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC Bioinformatics*, 18, 2017.

Ludivine Crible and Maria Josep Cuenca. Discourse markers in speech: characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8(2):149–166, 2017.

Dan Cristea, Nancy Ide, Daniel Marcu, and Valentin Tablan. Discourse structure and co-reference: An empirical study. In *The Relation of Discourse/Dialogue Structure and Reference*, 1999. URL `https://aclanthology.org/W99-0106`.

Zeyu Dai and Ruihong Huang. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2976–2987, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1295. URL `https://aclanthology.org/D19-1295`.

Parag Pravin Dakle, Takshak Desai, and Dan Moldovan. A study on entity resolution for email conversations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 65–73, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.8`.

Pascal Denis and Jason Baldridge. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 660–669, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL `https://aclanthology.org/D08-1069`.

Sobha Lalitha Devi. Anaphora Resolution from Social Media Text in Indian Languages (SocAnaRes-IL) Overview. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 9–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450389785. doi: 10.1145/3441501.3441512. URL `https://doi.org/10.1145/3441501.3441512`.

Joseph A. DeVito. Psychogrammatical factors in oral and written discourse by skilled communicators. *Speech Monographs*, 33(1):73–76, 1966.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Jelske Dijkstra, Wilbert Heeringa, Lysbeth Jongbloed-Faber, and Hans Van de Velde. Using twitter data for the study of language change in low-resource languages. a panel study of relative pronouns in frisian. *Frontiers in Artificial Intelligence*, 4, 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.644554. URL `https://www.frontiersin.org/article/10.3389/frai.2021.644554`.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. Adapting Coreference Resolution for Narrative Processing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2267, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1271. URL `https://www.aclweb.org/anthology/D15-1271`.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference* *on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

Gonzalo Donoso and David Sánchez. Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 16–25, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1202. URL https://aclanthology.org/W17-1202.

Greg Durrett and Dan Klein. Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1203.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1245. URL https://aclanthology.org/D17-1245.

Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1037.

S. Engell. *Coreference in English and German: A Theoretical Framework and Its Application in a Study of Court Decisions.* Logos Verlag Berlin, 2016. ISBN 9783832543396. URL https://books.google.de/books?id=sjGjAQAACAAJ.

Kelly Enochson and Jennifer Culbertson. Collecting psycholinguistic response time data using amazon mechanical turk. *PLOS ONE*, 10(3):1–17, 03 2015. doi: 10.1371/journal.pone.0116946. URL https://doi.org/10.1371/journal.pone.0116946.

B. Everitt, S. Landau, M. Leese, D. Stahl, and an O'Reilly Media Company Safari. *Cluster Analysis, 5th Edition.* John Wiley & Sons, 2011.

Mark Finlayson, Jeffry Halverson, and Steven Corman. The n2 corpus: A semantically annotated collection of islamist extremist stories. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 896–902, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/48_Paper.pdf.

Mark A. Finlayson and Raquel Hervás. Ucm/mit indications, referring expressions, and coreference corpus (umirec corpus) v1.1. In *MIT CSAIL Work Product*, 2010.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420, June 2011. doi: 10.1162/COLI_a_00057. URL https://aclanthology.org/J11-2010.

Barbara A. Fox. *Discourse structure and anaphora : written and conversational English / Barbara A. Fox.* Cambridge University Press Cambridge [Cambridgeshire] ; New York, 1987. ISBN 0521330823. URL http://www.loc.gov/catdir/toc/cam031/86033349.html.

Jean E. Fox Tree and Josef C. Schrock. Basic meanings of you know and i mean. *Journal of Pragmatics*, 34(6):727 – 747, 2002. ISSN 0378-2166. doi: https://doi.org/10.1016/S0378-2166(02)00027-9. URL `http://www.sciencedirect.com/science/article/pii/S0378216602000279`.

G. Frege. "uber sinn und bedeutung. In Mark Textor, editor, *Funktion - Begriff - Bedeutung*, volume 4 of *Sammlung Philosophie*. Vandenhoeck & Ruprecht, G"ottingen, 1892.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 2017.

S.C. Garrod and A. J. Sanford. THE MENTAL REPRESENTATION OF DISCOURSE IN A FOCUSSED MEMORY SYSTEM: IMPLICATIONS FOR THE INTERPRETATION OF ANAPHORIC NOUN PHRASES. *Journal of Semantics*, 1(1):21–41, 01 1982.

Abbas Ghaddar and Phillippe Langlais. Coreference in Wikipedia: Main concept resolution. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 229–238, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/K16-1023. URL `https://aclanthology.org/K16-1023`.

Abbas Ghaddar and Phillippe Langlais. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia, May 2016b. European Language Resources Association (ELRA). URL `https://aclanthology.org/L16-1021`.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6. URL `http://dl.acm.org/citation.cfm?id=2002736.2002747`.

Talmy Givón, editor. *Topic Continuity in Discourse: A Quantitative Cross-Language Study.* John Benjamins., Amsterdam, 1983.

J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development . acoustics,. In *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, volume 1, pages 517–520, 1992.

M. Greenacre. *Correspondence Analysis in Practice (3rd ed.).* Chapman and Hall/CRC, 2016. URL `https://doi.org/10.1201/9781315369983`.

Ralph Grishman and Beth Sundheim. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. URL `https://aclanthology.org/C96-1079`.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995. URL `https://aclanthology.org/J95-2003`.

Jeanette Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307, 1993.

Grigorii Guz and Giuseppe Carenini. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.codi-1.17. URL `https://aclanthology.org/2020.codi-1.17`.

M. A. K. Halliday. *Spoken and Written Language.* Deakin University Press, 1985.

M. A. K. Halliday and R. Hasan. *Cohesion in English.* Longman, London, 1976.

M.A.K. Halliday. *An introduction to functional grammar.* E. Arnold, London, 2004.

Boran Hao, Henghui Zhu, and Ioannis Paschalidis. Enhancing clinical BERT embedding using a biomedical knowledge base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.57. URL `https://aclanthology.org/2020.coling-main.57`.

Christian Hardmeier, Luca Bevacqua, Sharid Loáiciga, and Hannah Rohde. Forms of Anaphoric Reference to Organisational Named Entities: Hoping to widen appeal, they diversified. In *Proceedings of the Seventh Named Entities Workshop*, pages 36–40, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Lynette Hirschman and Nancy Chinchor. Appendix F: MUC-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998. URL `https://aclanthology.org/M98-1029`.

Jerry R. Hobbs. Pronoun resolution. Technical report, Department of Computer Sciences, City College, City University of New York., August 1976. Research Report 76-1.

Jerry R. Hobbs. Coherence and coreference. *Cognitive Science*, 3(1):67–90, 1979. ISSN 0364-0213. URL `https://www.sciencedirect.com/science/article/pii/S0364021379800439`.

J.R. Hobbs. Resolving pronoun references. *Lingua 44*, pages 311–338, 1978.

Jet Hoek, Merel C.J. Scholman, and Ted J.M. Sanders. Is there less agreement when the discourse is underspecified? In *Proceedings of the DiscAnn Workshop*, 2021.

Anke Holler and Lisa Irmen. Empirically assessing effects of the right frontier constraint. In António Branco, editor, *Anaphora: Analysis, Algorithms and Applications*, pages 15–27, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-71412-5.

Veronique Hoste. *Optimization issues in machine learning of coreference resolution.* PhD thesis, Antwerp University, 2005. URL `http://lib.ugent.be/fulltxt/RUG01/000/898/325/RUG01-000898325\_2010\_0001\_AC.pdf`.

Yufang Hou, Katja Markert, and Michael Strube. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284, June 2018. doi: 10.1162/COLI_a_00315. URL `https://aclanthology.org/J18-2002`.

Neslihan Iskender, Robin Schaefer, Tim Polzehl, and Sebastian Möller. Argument Mining in Tweets: Comparing Crowd and Expert Annotations for Automated Claim and Evidence Detection. In H. Horacek E. Métais, F. Meziane and E. Kapetanios, editors, *Natural Language Processing and Information Systems (NLDB)*, Lecture Notes in Computer

Science. Springer, Cham, June 2021. doi: https://doi.org/10.1007/978-3-030-80599-9_25.

Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188, 2009. doi: https://doi.org/10.1002/asi.21149. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21149`.

Ewa Jonsson. *Conversational Writing: A Multidimensional Study of Synchronous and Supersynchronous Computer-mediated Communication*. English Corpus Linguistics. Peter Lang, Frankfurt am Main, 2016. ISBN 9783631671535.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1588. URL `https://aclanthology.org/D19-1588`.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. doi: 10.1162/tacl_a_00300. URL `https://aclanthology.org/2020.tacl-1.5`.

Ben Kantor and Amir Globerson. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1066. URL `https://aclanthology.org/P19-1066`.

Andrew Kehler. *Coherence, reference and the theory of grammar*. Stanford: CSLI Publications, 2002.

Andrew Kehler and Hannah Rohde. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39 (1-2):1–37, 2013. doi: doi:10.1515/tl-2013-0001. URL `https://doi.org/10.1515/tl-2013-0001`.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. Coherence and coreference revisited. *Journal of Semantics*, 25(1):1–44, 2008.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.codi-sharedtask.1. URL `https://aclanthology.org/2021.codi-sharedtask.1`.

Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1249. URL `https://aclanthology.org/P18-1249`.

Wolfgang Klein and Christiane Stutterheim. Text structure and referential movement. *Zeitschrift für Literaturwissenschaft und Linguistik*, 22(86):67–92, 1992.

P. Koch and W. Oesterreicher. Language of immediacy – language of distance: Orality and literacy from the perspective of language theory and linguistic history. In B. Weber C. Lange and G. Wolf, editors, *Communicative Spaces. Variation, Contact, and Change: Papers in Honour of Ursula Schaefer*. Peter Lang, Frankfurt am Main, 2012.

Peter Koch and Wulf Oesterreicher. Sprache der Nähe - Sprache der Distanz. Müdlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(1985):15 – 43, 1985. doi: https://doi.org/10.1515/9783110244922.15. URL https://www.degruyter.com/view/journals/roma/36/1985/article-p15.xml.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. Anaphora With Non-nominal Antecedents in Computational Linguistics: a Survey. *Computational Linguistics*, 44(3):547–612, 09 2018. ISSN 0891-2017. doi: 10.1162/coli_a_00327. URL https://doi.org/10.1162/coli_a_00327.

Mateusz Kopec and Maciej Ogrodniczuk. Inter-annotator agreement in coreference annotation of polish. In J. Sobecki, V. Boonjing, and S. Chittayasothorn, editors, *Advanced Approaches to Intelligent Information and Database Systems, Studies in Computational Intelligence*, volume 551. Switzerland: Springer. Springer International Publishing, Switzerland, 2014.

K. Krippendorff. *Content Analysis: An Introduction To Its Methodology*. Sage commtext series. Sage Publications, 1980.

Kerstin Kunz. A method for investigating coreference in translations and originals. *Languages in Contrast*, 7(2):267–287, 2007. ISSN 1387-6759. doi: https://doi.org/10.1075/lic.7.2.10kun. URL https://www.jbe-platform.com/content/journals/10.1075/lic.7.2.10kun.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. Cross-linguistic analysis of discourse variation across registers. *Cross-linguistic Studies at the Interface between Lexis and Grammar. Nordic Journal of English Studies.*, 14:258–288, 2015. URL http://ojs.ub.gu.se/ojs/index.php/njes/article/view/3095.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, and José Manuel Martínez. Beyond identity coreference: Contrasting indicators of textual coherence in english and german. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 23–31. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-0704. URL http://www.aclweb.org/anthology/W16-0704.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, José Manuel Martínez Martínez, Katrin Menzel, and Erich Steiner. *GECCo - German-English Contrasts in Cohesion*. De Gruyter Mouton, Berlin, Boston, 2021. ISBN 9783110711073. doi: doi:10.1515/9783110711073. URL https://doi.org/10.1515/9783110711073.

W. Labov and J. Waletzky. Narrative Analysis. In J. Helm, editor, *Essays on the Verbal and Visual Arts*, pages 12–44. U. of Washington Press, 1967.

Robin Tolmach Lakoff. Some of My Favorite Writers are Literate: The Mingling of Oral and Literate Strategies in Written Communication. In *Spoken and Written Language: Exploring Orality and Literacy*. ABLEX Pub. Corp, Norwood, N.J, 1982.

Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994. URL https://aclanthology.org/J94-4002.

Ekaterina Lapshinova-Koltunski. Exploration of inter- and intralingual variation of discourse phenomena. In *Proceedings of the Second Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2015, Lisbon, Portugal, September 17, 2015*, pages 158–167, 2015. doi: 10.18653/v1/W15-2521. URL `https://doi.org/10.18653/v1/W15-2521`.

Ekaterina Lapshinova-Koltunski and Kerstin Kunz. Exploring coreference features in heterogeneous data. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 53–64, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.codi-1.6. URL `https://aclanthology.org/2020.codi-1.6`.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1065`.

Ekaterina Lapshinova-Koltunski, Marie-Pauline Krielke, and Christian Hardmeier. Coreference strategies in English-German translation. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 139–153, Barcelona, Spain (online), December 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.crac-1.15`.

Tamar Lavee, Lili Kotlerman, Matan Orbach, Yonatan Bilu, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Crowd-sourcing annotation of complex NLU tasks: A case study of argumentative content annotation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 29–38, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5905. URL `https://aclanthology.org/D19-5905`.

Alan Lee, Rashmi Prasad, Bonnie Webber, and Aravind K. Joshi. Annotating discourse relations with the PDTB annotator. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 121–125, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://aclanthology.org/C16-2026`.

David Yong Wey Lee. Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning and Technology*, 5:37–72, 2001.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284084. URL `http://dl.acm.org/citation.cfm?id=2132936.2132938`.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, 39(4):885–916, 12 2013. ISSN 0891-2017. doi: 10.1162/COLI_a_00152. URL `https://doi.org/10.1162/COLI_a_00152`.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in*

Natural Language Processing, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL `https://www.aclweb.org/anthology/D17-1018`.

Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2108. URL `https://www.aclweb.org/anthology/N18-2108`.

Sven Leuckert and Sarah Buschfeld. Modelling spoken and written language. *Anglistik*, 32(2), 2021. doi: 10.33675/ANGL/2021/2/4.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press, Rome, Italy, 2012. URL `https://cs.nyu.edu/faculty/davise/papers/WSKR2012.pdf`.

Xuansong Li, Martha Palmer, Nianwen Xue, Lance Ramshaw, Mohamed Maamouri, Ann Bies, Kathryn Conger, Stephen Grimes, and Stephanie Strassel. Large multi-lingual, multi-level and multi-genre annotation corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 906–913, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL `https://aclanthology.org/L16-1145`.

Yufei Li, Xiaoyong Ma, Xiangyu Zhou, Pengzhen Cheng, Kai He, and Chen Li. Knowledge enhanced LSTM for coreference resolution on biomedical texts. *Bioinformatics*, 37(17): 2699–2705, 03 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab153. URL `https://doi.org/10.1093/bioinformatics/btab153`.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-2013`.

Pengcheng Lu and Massimo Poesio. Coreference resolution for the biomedical domain: A survey. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 12–23, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.crac-1.2. URL `https://aclanthology.org/2021.crac-1.2`.

Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL `https://aclanthology.org/H05-1004`.

Xiaoqiang Luo, Abraham Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 135–142, 01 2004. doi: 10.3115/1218955.1218973.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. An extension of BLANC to system mentions. In *Proceedings of the 52nd Annual Meeting of the*

Association for Computational Linguistics (Volume 2: Short Papers), pages 24–29, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2005. URL `https://aclanthology.org/P14-2005`.

William Mann and Sandra Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988. doi: doi:10.1515/text.1.1988.8.3.243. URL `https://doi.org/10.1515/text.1.1988.8.3.243`.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-5010. URL `https://aclanthology.org/P14-5010`.

J. R. Martin. *Factual writing : exploring and challenging social reality.* Deakin University Press, 1985.

James R. Martin. Language, register and genre. In A. Burns and C. Coffin, editors, *Analysing English in a Global Context: a reader*, Teaching English Language Worldwide, pages 149–166. Routledge, Clevedon, 2001. Revised version of [Martin, 1984].

Ruslan Mitkov. *Anaphora Resolution.* Pearson Education, 2002.

Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1060. URL `https://aclanthology.org/P16-1060`.

Nafise Sadat Moosavi and Michael Strube. Using linguistic features to improve the generalization capability of neural coreference resolvers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1018. URL `https://aclanthology.org/D18-1018`.

Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with mmax2. In Joybrato Mukherjee Sabine Braun, Kurt Kohn, editor, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, 2006.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1001. URL `https://aclanthology.org/S16-1001`.

Gerald Nelson, Sean Wallis, and Bas Aarts. *Exploring Natural Language: Working with the British Component of the International Corpus of English.* Varieties of English Around the World S. John Benjamins Publishing Company, Philadelphia, 2002. ISBN 9789027248886.

Stella Neumann and Jennifer Fest. Cohesive devices across registers and varieties: The role of medium in English. In *Variational text linguistics : revisiting register in English / edited by Christoph Schubert, Christina Sanchez-Stockhammer*, volume 90 of *Topics* *in English Linguistics*, pages 195–220. De Gruyter, Berlin, Boston, 2016. doi: 10.1515/9783110443554-010. URL https://publications.rwth-aachen.de/record/658390.

Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1142.

Vincent Ng. Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4877–4884. AAAI Press, 2017.

Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 104–111, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073102. URL https://doi.org/10.3115/1073083.1073102.

Ngan Nguyen, Jin-Dong Kim, and Jun'ichi Tsujii. Overview of BioNLP 2011 protein coreference shared task. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 74–82, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-1811.

Elinor Ochs. *Planned and Unplanned Discourse*, pages 51 – 80. Brill, Leiden, The Netherlands, 1979. ISBN 9789004368897. doi: https://doi.org/10.1163/9789004368897_004. URL https://brill.com/view/book/edcoll/9789004368897/BP000004.xml.

Maciej Ogrodniczuk and Vincent Ng, editors. *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-07. URL https://aclanthology.org/W16-0700.

Maciej Ogrodniczuk and Vincent Ng, editors. *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-15. URL https://aclanthology.org/W17-1500.

Maciej Ogrodniczuk, Vincent Ng, Yulia Grishina, and Sameer Pradhan, editors. *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, Barcelona, Spain (online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.crac-1.0.

Maciej Ogrodniczuk, Sameer Pradhan, Massimo Poesio, Yulia Grishina, and Vincent Ng, editors. *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.crac-1.0.

Niki Panteli, Andriana Rapti, and Dora Scholarios. 'If He Just Knew Who We Were': Microworkers' Emerging Bonds of Attachment in a Fragmented Employment Relationship. *Work, Employment and Society*, 34(3):476–494, 2020. doi: 10.1177/0950017019897872.

Rebecca Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

Rebecca J. Passonneau. Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/752.pdf`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://aclanthology.org/N18-1202`.

Saša Petrović, Miles Osborne, and Victor Lavrenko. The Edinburgh Twitter Corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, page 25–26, USA, 2010. Association for Computational Linguistics.

Arianna Pipitone, Giuseppe Tirone, and Roberto Pirrone. Named entity recognition and linking in tweets based on linguistic similarity. In Floriana Esposito, Roberto Basili, Stefano Ferilli, and Francesca A. Lisi, editors, *AI*IA 2017 Advances in Artificial Intelligence*, pages 101–113, Cham, 2017. Springer International Publishing. ISBN 978-3-319-70169-1.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0702. URL `https://aclanthology.org/W18-0702`.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1176. URL `https://aclanthology.org/N19-1176`.

Livia Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5): 601–638, 1988. ISSN 0378-2166. doi: https://doi.org/10.1016/0378-2166(88)90050-1. URL `https://www.sciencedirect.com/science/article/pii/0378216688900501`.

Simone Paolo Ponzetto and Michael Strube. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 192–199, New

York City, USA, June 2006. Association for Computational Linguistics. URL https://aclanthology.org/N06-1025.

Marta Recasens Potau. Coreference: Theory, annotation, resolution and evaluation. *Unpublished Dissertation. Ph. D Program. University of Barcelona*, 2010.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica Macbride, and Linnea Micciulla. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. *International Conference on Semantic Computing*, 0:446–453, 09 2007. doi: 10.1109/ICSC.2007.93.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40. Association for Computational Linguistics, 2012.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152. Association for Computational Linguistics, 2013.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-2006.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The Penn Discourse Treebank 2.0. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

Rashmi Prasad, Bonnie Webber, and Alan Lee. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

Ellen F. Prince. The zpg letter: Subjects, definiteness, and information-status. In William C. Mann and Sandra A. Thompson, editors, *Discourse Description: Diverse linguistic analyses of a fund-raising text*, 1992.

Thomas Proisl and Peter Uhrig. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin, 2016. Association for Computational Linguistics (ACL). URL http://aclweb.org/anthology/W16-2607.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14. URL https://aclanthology.org/2020.acl-demos.14.

Altaf Rahman and Vincent Ng. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/D09-1101.

M. Recasens and E. Hovy. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510, oct 2011. ISSN 1351-3249. doi: 10.1017/S135132491000029X. URL `https://doi.org/10.1017/S135132491000029X`.

Ines Rehbein, Merel Scholman, and Vera Demberg. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016.

Giuseppe Riccardi, Evgeny A. Stepanov, and Shammur Absar Chowdhury. Discourse connective detection in spoken conversations. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6095–6099, 2016. doi: 10.1109/ICASSP.2016.7472848.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `https://aclanthology.org/D11-1141`.

Ina Roesiger. Rule- and learning-based methods for bridging resolution in the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 23–33, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0703. URL `https://aclanthology.org/W18-0703`.

Ina Roesiger, Arndt Riester, and Jonas Kuhn. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://aclanthology.org/C18-1298`.

Patricia Ronan. Tweeting with Trump. *Anglistik*, 32:67–83, 01 2021. doi: 10.33675/ANGL/2021/2/7.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL `https://aclanthology.org/N18-2002`.

Attapol Rutherford and Nianwen Xue. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1068. URL `https://aclanthology.org/E14-1068`.

Sofia Rüdiger. Digital food talk: Blurring immediacy and distance in youtube eating shows. *Anglistik*, 32(2), 2021. doi: 10.33675/ANGL/2021/2/9. URL `https://doi.org/10.33675/ANGL/2021/2/9`.

Harvey Sacks, Emanuel A. Schegloff, and Gail D. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696 – 735, 1974.

Fahime Same and Kees van Deemter. A linguistic perspective on reference: Choosing a feature set for generating referring expressions in context. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4575–4586, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.403. URL `https://aclanthology.org/2020.coling-main.403`.

Ted Sanders, Wilbert Spooren, and Leo G. M. Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35, 1992.

Robin Schaefer and Manfred Stede. Argument mining on Twitter: A survey. *it - Information Technology*, 63(1):45–58, 2021. doi: doi:10.1515/itit-2020-0053. URL `https://doi.org/10.1515/itit-2020-0053`.

Ursula Schaefer. Communicative Distance: The (Non-)Reception of Koch and Oesterreicher in English-Speaking Linguistics. *Anglistik*, 32(2), 2021. doi: 10.33675/ANGL/2021/2/5. URL `https://doi.org/10.33675/ANGL/2021/2/5`.

Tatjana Scheffler. A German Twitter snapshot. In *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 2284–2289, Reykjavik, Iceland, 2014.

Tatjana Scheffler. Conversations on Twitter. In Darja Fišer and Michael Beißwenger, editors, *Researching computer-mediated communication: Corpus-based approaches to language in the digital world*, pages 124–144. University Press, Ljubljana, 2017.

Tatjana Scheffler and Manfred Stede. Realizing argumentative coherence relations in German: A contrastive study of newspaper editorials and Twitter posts. In *Proceedings of the COMMA Workshop "Foundations of the Language of Argumentation"*, Potsdam, Germany, 2016.

Catherine Schnedecker. Reference chains and genre identification: From Discrete to Non-Discrete Units. In Thierry Charnois Dominique Legallois and Meri Larjavaara, editors, *The Grammar of Genres and Styles*, page 39–66. De Gruyter Mouton, 2018. doi: 10.1515/9783110595864-003.

Uta Schäpers. *Nominal versus Clausal Complexity in Spoken and Written English.* Peter Lang Verlag, 2009.

Sandhya Singh, Kevin Patel, and Pushpak Bhattacharyya. Attention based anaphora resolution for code-mixed social media text for hindi language. In Parth Mehta, Thomas Mandl, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020*, volume 2826 of *CEUR Workshop Proceedings*, pages 780–787. CEUR-WS.org, 2020. URL `http://ceur-ws.org/Vol-2826/T8-1.pdf`.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001. doi: 10.1162/089120101753342653. URL `https://aclanthology.org/J01-4004`.

Robyn Speer and Catherine Havasi. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/1072_Paper.pdf`.

Manfred Stede. *Discourse Processing.* Morgan and Claypool Publishers, 2011. ISBN 1608457346.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Ježek. Two uses of anaphora resolution in summarization. *Information Processing & Management*, 43(6): 1663–1680, 2007. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2007.01.010. URL `https://www.sciencedirect.com/science/article/pii/S0306457307000428`. Text Summarization.

J. Streb, E. Hennighausen, and F. Rösler. Different anaphoric expressions are investigated by event-related brain potentials. *Journal of Psycholinguistic Research*, 33:175–201, 2004.

Michael Strube and Udo Hahn. Functional centering. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, page 270–277, USA, 1996. Association for Computational Linguistics. doi: 10.3115/981863.981899. URL `https://doi.org/10.3115/981863.981899`.

Nikolaos Stylianou and Ioannis Vlahavas. A neural entity coreference resolution review. *Expert Systems with Applications*, 168:114466, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.114466. URL `https://www.sciencedirect.com/science/article/pii/S0957417420311143`.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.01.010. URL `https://www.sciencedirect.com/science/article/pii/S1566253519303677`.

W. Swanson. *Modes of Co-reference as an Indicator of Genre.* Linguistic insights. P. Lang, 2003. ISBN 9780820468556. URL `https://books.google.de/books?id=OrwaPwAACAAJ`.

D. Tannen. The oral/literate continuum in discourse. In D. Tannen, editor, *Spoken and Written Language: Exploring Orality and Literacy*, page 1–16. Ablex, Norwood, NJ, 1982a.

Deborah Tannen. Oral and literate strategies in spoken and written narratives. *Language*, 58:1–21, 1982b.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. The Penn Treebank: An Overview. In Abeillé A., editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology.* Springer, Dordrecht, 2003.

Joel R. Tetreault. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520, 2001. URL `https://aclanthology.org/J01-4003`.

Hariprasad Timmapathini, Anmol Nayak, Sarathchandra Mandadi, Siva Sangada, Vaibhav Kesri, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. Probing the SpanBERT Architecture to interpret Scientific Domain Adaptation Challenges for Coreference Resolution. In Amir Pouran Ben Veyseh, Franck Dernoncourt, Thien Huu Nguyen, Walter Chang, and Leo Anthony Celi, editors, *Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021*, volume 2831 of *CEUR Workshop Proceedings.* CEUR-WS.org, 2021. URL `http://ceur-ws.org/Vol-2831/paper10.pdf`.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/184_Paper.pdf`.

Janine Toole. The effect of genre on referential choice. In Thorstein Fretheim and Jeanette K. Gundel, editors, *Reference and Referent Accessibility*, pages 263–290. John Benjamins, Amsterdam, 1996.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.685. URL `https://aclanthology.org/2020.emnlp-main.685`.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.crac-1.12. URL `https://aclanthology.org/2021.crac-1.12`.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1062. URL `https://aclanthology.org/D19-1062`.

Olga Uryupina and Alessandro Moschitti. A state-of-the-art mention-pair model for coreference resolution. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 289–298, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-1034. URL `https://aclanthology.org/S15-1034`.

Olga Uryupina and Massimo Poesio. Domain-specific vs. Uniform Modeling for Coreference Resolution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 187–191, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/944_Paper.pdf`.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Kepa Rodriguez, and Massimo Poesio. ARRAU: Linguistically-motivated annotation of anaphoric descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2058–2062, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL `https://aclanthology.org/L16-1326`.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995. URL `https://aclanthology.org/M95-1005`.

Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Commun. ACM*, 51 (8):58–67, 2008. ISSN 0001-0782.

Klaus von Heusinger and Petra B. Schumacher. Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117–127, 2019. ISSN 0378-2166. doi: https://doi.org/10.1016/j.pragma.2019.07.025.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1566. URL `https://aclanthology.org/P19-1566`.

Thomas Wasow. Remarks on grammatical weight. *Language Variation and Change*, 9:81 – 105, 03 1997. doi: 10.1017/S0954394500001800.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. The Penn Discourse Treebank 3.0 Annotation Manual. Report, The University of Pennsylvania, 2018.

Bonnie Lynn Webber. Discourse deixis and discourse processing. Technical report, University of Pennsylvania, 1988.

Juliet Webster. Microworkers of the gig economy: Separate and precarious. *New Labor Forum*, 25(3):56–64, 2016. doi: 10.1177/1095796016661511.

Zhao Wei. A survey of studies of bridging anaphora. *Canadian Social Science*, 10:130–139, 2014.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. Ontonotes release 5.0 ldc2013t19. *Web Download. Linguistic Data Consortium, Philadelphia, PA*, 2013.

Terry Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(72)90002-3. URL `https://www.sciencedirect.com/science/article/pii/0010028572900023`.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1137. URL `https://aclanthology.org/P15-1137`.

C. H. Woolbert. Speaking and writing: A study of differences. *Quarterly Journal of Speech*, 8(3):271–285, 1922. doi: 10.1080/00335632209379390.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.622. URL `https://aclanthology.org/2020.acl-main.622`.

Patrick Xia and Benjamin Van Durme. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing, pages 5241–5256, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.425. URL `https://aclanthology.org/2021.emnlp-main.425`.

Patrick Xia, João Sedoc, and Benjamin Van Durme. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.695. URL `https://aclanthology.org/2020.emnlp-main.695`.

Liyan Xu and Jinho D. Choi. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.686. URL `https://aclanthology.org/2020.emnlp-main.686`.

Liyan Xu and Jinho D. Choi. Adapted end-to-end coreference resolution system for anaphoric identities in dialogues. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 55–62, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.codi-sharedtask.6. URL `https://aclanthology.org/2021.codi-sharedtask.6`.

Wei Xu. From shakespeare to Twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation*, pages 1–9, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4901. URL `https://aclanthology.org/W17-4901`.

Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, Ting Liu, and Sheng Li. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of ACL-08: HLT*, pages 843–851, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `https://aclanthology.org/P08-1096`.

Juntao Yu, Alexandra Uma, and Massimo Poesio. A cluster ranking model for full anaphora resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 11–20, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.2`.

Nan Yu, Meishan Zhang, and Guohong Fu. Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://aclanthology.org/C18-1047`.

Amir Zeldes. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017. doi: http://dx.doi.org/10.1007/s10579-016-9343-x.

Amir Zeldes. A Predictive Model for Notional Anaphora in English. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 34–43, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Deniz Zeyrek and Murathan Kurfalı. TDB 1.1: Extensions on Turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-0809. URL `https://aclanthology.org/W17-0809`.

Linrui Zhang, Yisheng Zhou, Yang Yu, and Dan I. Moldovan. Towards understanding creative language in tweets. *Journal of Software Engineering and Applications*, 2019.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. Anatomy of OntoGUM—Adapting GUM to the OntoNotes scheme to evaluate robustness of SOTA coreference algorithms. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 141–149, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.crac-1.15. URL `https://aclanthology.org/2021.crac-1.15`.

Šárka Zikánová, Jiří Mírovský, and Pavlína Synková. Explicit and implicit discourse relations in the prague discourse treebank. In Kamil Ekštein, editor, *Text, Speech, and Dialogue*, pages 236–248, Cham, 2019. Springer International Publishing.

G. K. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Reading MA (USA), 1949.