Variation in Coreference Strategies across Genres and Production Media

Berfin Aktaş
SFB1287
Research Focus Cognitive Sciences
University of Potsdam, Germany
berfinaktas@uni-potsdam.de

Manfred Stede
SFB1287
Research Focus Cognitive Sciences
University of Potsdam, Germany
stede@uni-potsdam.de

Abstract

In response to (i) inconclusive results in the literature as to the properties of coreference chains in written versus spoken language, and (ii) a general lack of work on automatic coreference resolution on both spoken language and social media, we undertake a corpus study involving the various genre sections of Ontonotes, the Switchboard corpus, and a corpus of Twitter conversations. Using a set of measures that previously have been applied individually to different data sets, we find fairly clear patterns of ”behavior” for the different genres/media. Besides their role for psycholinguistic investigation (why do we employ different coreference strategies when we write or speak) and for the placement of Twitter in the spoken–written continuum, we see our results as a contribution to approaching genre-/media-specific coreference resolution.

1 Introduction

Research on strategies for producing referring expressions has often investigated the differences (if any) between spoken and written language, but as we will show in Section 2, results have been inconclusive. Sometimes, claims are simply contradictory, but the more important problem is that usually, the exact ways of measuring the properties of coreference chains are not being made transparent. In addition, the data that has been used can vary considerably, and it is not always clear how studies can be compared.

Our primary goal here is to shed light on coreference with respect to the spoken/written distinction, by undertaking a careful comparative corpus analysis and explicitly stating our methods of measurement. In an earlier study, we presented a quantitative study on different genre sections of the OntoNotes corpus (Aktaş et al., 2019). We extend that study in two directions: We augment the rather small proportion of spoken data in OntoNotes with the Switchboard corpus (Godfrey et al., 1992), and we introduce new quantitative features for the comparative analysis.

The secondary goal is to explore how the medium *microblog*, specifically Twitter, relates to the spoken-written spectrum for coreference strategies. In a recent study, we investigated this research question empirically through computational experiments on a state-of-the-art ”standard” coreference resolution system (Aktaş et al., 2020). We showed that the choice of genre and the medium (spoken vs. written) in training data can make a bigger difference than the bare amount of data. In the current paper, we apply corpus-based empirical methods to situate the coreference strategies found in Twitter conversations in the medium/genre spectrum. For this purpose, we finally extend the (Aktaş et al., 2019) study by adding the comparison with Twitter texts, thus achieving a wider range of production media.

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.

We will show genre-specific distributional patterns of nominal referring expressions in terms of **frequency** of syntactic categories, **heaviness** of NP structures, and relative **distance** between anaphors and antecedents in the text. Most of our analyses lead to a common ranking and clustering of the genres based on the measures and results presented in Section 4, as will be discussed in Section 5. In addition to the theoretical value, we believe that the observed patterns can provide a basis for data-driven design of automated coreference resolution tools that perform better on spoken or Twitter language than the current ”out of the box” systems do.

The following section presents related work. Section 3 introduces data sources and corpus alignment strategies we applied to harmonize the data. Section 4 presents the measures we used for the comparison, and the results of the analyses, which are further discussed in Section 5, concentrating on the comparison of the genres and production media. Section 6 summarizes and shows directions for future work.

2 Related Work

Coreference across genres and media. Various linguistic coreference phenomena have been compared by researchers in different domains such as across languages (Lapshinova-Koltunski, 2015; Kunz and Lapshinova-Koltunski, 2015; Engell, 2016; Kunz et al., 2016), regional language varieties (Neumann and Fest, 2016), production media (spoken, written, web) (Fox, 1987; Biber, 1992; Amoia et al., 2012) and across genres in these domains. Among the features, frequency-based statistics and distance measurements are 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, and the results sometimes point into different directions. For example, Biber et al. (1999) and Kunz et al. (2016) measured the distance in terms of number of tokens. Both teams found that the average distance is longer in spoken texts than in written texts. Fox (1987) measured the distance in terms of number of clauses, and arrived at the same observation. In contrast, Amoia et al. (2012) measured the distance in terms of sentences and concluded that the average distance is longer in written texts than in spoken texts. The same claim was made by Biber (1992), who computed distance as the number of interfering mentions. These partly-incompatible outcomes indicate that textual distance metrics are not easily comparable; for instance, the distance in terms of tokens may not always correspond to distance in terms of clauses or sentences. This is one of the aspects we will address in this paper.

Other 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) argued 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) found different characteristics: On the one hand, they confirm Fox’s finding that NPs are more frequently used than pronouns in the written medium than in the spoken medium. But in written text, 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). For length, they found that the avg. number of tokens of referring expressions in their written data is 3.42, compared to 2.58 for spoken data. Another interesting finding on the distribution of syntactic categories is that narrative genres (fiction) show spoken-like characteristics in terms of pronoun usage (Biber, 1992; Biber et al., 1999; Amoia et al., 2012; Neumann and Fest, 2016; Lapshinova-Koltunski, 2015).

Further quantitative metrics of coreference phenomena used in the literature are the number of referring expressions (Biber, 1992; Schnedecker, 2018), 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 did not examine these features because in OntoNotes, documents are artificially split in smaller parts, and because singletons are not annotated (see Section 3); and in Switchboard, there exist unannotated coreference chains in the data (see Section 3.4). 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 investigated how organisational entities are being referred to, and found a correlation between preferred reference type and genre (e.g., pronouns are more common in telephone conversations than newswire and broadcast news). Zeldes (2018) used OntoNotes for predicting ’notional anaphora’ (i.e. pronouns disagreeing with their antecedents’ grammatical categories for notional reasons, e.g., “the government ... they.”), and found it to be more common in broadcast conversations than in newswire. Zeldes used 20 linguistic features, and genre emerged as the third-most important one, indicating that differences between genres can have an impact on automatic classification.

Automatic coreference resolution for "non-standard" language. Only few studies have addressed performance differences of coreference systems across the genres of Ontonotes. Pradhan et al. (2013) reported that their system of choice showed better performance on telephone conversations (64%) than on news texts (56%) and broadcast news (59%). The authors evaluated which sections turned out to be the "easiest" but did not assess the possible reasons. Another study that looks in detail at performance differences in OntoNotes (and also in two other corpora) is that of Uryupina and Poesio (2012), who compared the performance of domain-specific and generic models, for both knowledge-poor statistical systems and for implementations using hand-crafted linguistic features.

Similarly, coreference resolution specifically for spoken data has not received much attention. Beside work on the rather domain-specific data of the TRAINS corpus by Byron (2002), Eckert and Strube (2000) worked with the Switchboard corpus and gave results for an algorithm resolving personal and demonstrative pronouns. They found that pronoun usage differs considerably from that in written text, because of a high frequency of both non-nominal antecedents and pronouns with no antecedent at all. Strube and Müller (2003) refined the algorithm, and we are not aware of more recent work in this vein.

While Twitter has often been tackled for standard preprocessing tasks like POS tagging or NER, to our knowledge there are no previous studies investigating coreference resolution on Twitter data except for our own studies (Aktaş et al., 2020; Aktaş et al., 2018). In (Aktaş et al., 2020), we report that a state-of-the-art coreference resolver trained on OntoNotes data performs worse on Twitter data (F1=45.18 on Twitter vs F1=72.6 on original test data). In the same study, we conducted experiments with different training data settings and improved the performance of the tested system on the Twitter data by 21.6%. In follow up work, we aim to further improve coreference resolution on the genre of Twitter conversations by customizing the resolver according to the genre-based differences, including the contrasts that emerged from the comparative analysis presented in the current paper.

3 Corpora

We use three English-language data sources for our empirical study. See Table 1 for a summary, whose numbers are computed after harmonizing the corpora with respect to the criteria described in Section 3.4.

	tw	**swbd**	**OntoNotes**					
Feature			**total**	**tc**	**bc**	**bn**	**nw**	**wb**
# of documents	185	147	2040	142	274	947	597	80
# of tokens	43K	248K	903K	103K	147K	225K	355K	71K
# of sentences	3503	30K	55K	14K	10K	12K	14K	3677
# of clauses	6719	41K	110K	18K	21K	27K	35K	8072
# of coreference chains	1525	6863	25K	2461	4518	8042	9328	1523
# of mentions	6352	23K	103K	15K	20K	28K	34K	5827
avg. doc. length (token)	233.5	1688.6	442.9	729.5	537.0	238.3	595.7	893.3
avg. sent. length (token)	12.3	8.1	16.3	7.3	13.6	18.6	24.1	19.4
avg. sent. length (clause)	1.9	1.9	2.1	1.7	2.1	2.3	2.4	2.2
avg. 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

Table 1: General statistics on the corpora (**tw**: Twitter conversations, **swbd**: Switchboard telephone conversations, **tc**: telephone conversations, **bc**: broadcast conversations, **bn**: broadcast news, **nw**: newswire, **wb**: web blogs)

3.1 TwiConv

The conversational Twitter corpus[1] (**tw**) was built by Aktaş et al. (2018). After constructing the conversational full tree structures for randomly chosen tweets that generated replies, we kept only the

[1] Scripts and data to reproduce the corpus can be found at https://github.com/berfingit/coreference-variation

longest thread (a path from the root to a leaf node) from each tree and discarded all other branches, so there is no tweet overlap in the data. The corpus holds 43K tokens distributed across 1756 tweets arranged in 185 threads. The sentence segments in tweets are identified using the SoMaJo sentence splitter (Proisl and Uhrig, 2016)[2]. Boundary errors in complicated cases (e.g., when sentences in the same tweet start with a lowercase letter or a hashtag, or users omit punctuation) were manually corrected. As shown in Table 1, 1525 coreference chains are annotated in the corpus. Of these, 1055 contain inter-tweet references (i.e., at least one of the mentions in the chain is located in a different tweet than the rest). Further investigation of the links for each anaphor-antecedent pair showed that 50% of the coreferential links are established across tweets. Hence, for coreference resolution it is indeed important to consider the conversation context.

3.2 OntoNotes

The OntoNotes corpus (**ont**) consists of multi-language data from a range of different sources, including translations, and offers gold annotations at different linguistic layers such as part of speech tags, syntactic constituent parses and coreference chains. In our study, we used only the original English data, in order to avoid effects from potential translation divergences. The resulting data portion consists of both spoken and written language. Spoken data includes telephone conversations (**tc**), and broadcast TV conversations (**bc**), whereas written data is composed of newswire texts (**nw**) and web blogs (**wb**). Finally there are broadcast news (**bn**), which are produced in the spoken medium but mostly contain prepared speeches (i.e., edited language).

We used the CoNLL-formatted OntoNotes data (Pradhan et al., 2013). As shown in Table 1, the corpus contains 903K tokens distributed across 2040 *documents*[3]. The distribution of the various genres is also presented in the table.

3.3 Switchboard

Switchboard (**swbd**) is a long-standing corpus of conversational speech (Godfrey et al., 1992). The original dataset 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 predefined list and expected to have a conversation on it. Calhoun et al. (2010) brought together the various annotations made on the corpus and delivered a combined resource in NITE XML (NXT) format. 147 of the dialogs in NXT-Switchboard, all in separate documents, are annotated for coreference. In Switchboard, coreference links are marked as anaphor-antecedent pairs. We constructed complete chains from these pairs. The corpus contains 248K *tokens* in 147 documents, where 23K mentions are annotated in 6863 coreference chains (see Table 1).

3.4 Corpus Homogenity

Transcription The spoken texts can differ in terms of the transcription procedures applied. For instance, in Switchboard, silence moments and spots of grammatical ellipsis are inserted as separate tokens into the transcribed texts, whereas in Ontonotes, only the surface linguistic forms and punctuations are considered as tokens. In addition to this, unlike OntoNotes, repairs and false starts (i.e. reparanda) are also included in Switchboard transcriptions. We do not take these additional tokens in Switchboard into consideration in this study. 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.

Tokenization All the investigated corpora follow PTB tokenization conventions[4] with various adaptations based on the specific string types included in the texts. For instance, smileys, 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 in the standard PTB tokenization scheme. The

[2] Although only German is mentioned as target language in the cited paper, SoMaJo works also on English web and social media texts.

[3] In OntoNotes terminology, documents are the units of independent annotation.

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

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 Annotation 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 come with 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, and therefore we manually checked and corrected the structures computed through these parses in our analysis. The applied procedures are described below.

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 type of mentions in the **tw** corpus.

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 excluded these non-nominal mentions in OntoNotes in our analyses.

Another difference in the coreference annotation schemes is encountered in Switchboard, where only markables with information status "old" are annotated for coreference (Calhoun et al., 2010). However, not all candidates of referring expressions compatible with the markable definition are annotated for information status. This indicates that annotated coreference chains do not cover the complete set of non-singleton entities in the texts 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 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 4.4).

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]

(1) [Parents]$_\mathbf{i}$ should be involved with [their]$_\mathbf{i}$ children's education. [..] If [parents]$_\mathrm{k}$ are dissatisfied with a school, [they]$_\mathrm{k}$ should have the option of switching to another. (OntoNotes, two chains)

(2) 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)

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

Another potential source of incompatibility is our treatment of internet language elements in TwiConv. We do not split the strings such as 'idk' (i.e. "I don't know") or '#findHer' into their sub-units, which can potentially cause missing eligible markables in the annotations. However, we harmonized most of the data, and therefore, do not expect the remaining incompatibilities to affect the computations significantly.

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

4 Data Analysis

We use frequency-, heaviness- and distance-metrics in the quantitative comparison of the (sub-)corpora surveyed in Table 1.

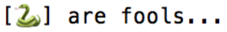
4.1 Delimiting NPs

For computing the frequency and heaviness measures we need to identify the boundaries of NPs in the texts. As mentioned in Section 3, OntoNotes and Switchboard have gold part-of-speech (PoS) and syntax annotations compatible to PTB conventions. These annotation layers are used to detect the NP structures in **swbd** and **ont**. 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. To detect the large 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 (e.g. ”an invasion of the privacy”). For short NP spans, we traverse the tree in a top down **depth-first** manner, but extract the **last** encountered NP nodes this time in each branch (e.g. ”an invasion” and ”the privacy”).

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

Regarding the non-standard tokens in Twitter, we consider emojis, smileys and links as NPs only if they have a syntactic role of an NP in a sentence as shown in brackets in examples (4) and (5) below. There are 4 cases among 553 emojis and 4 cases among 340 links in **tw** corpus, where emojis and links are considered as NPs.

(4) 

[🐍] are fools...

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

The automatically-inserted usernames at the beginning of reply tweets are not considered as NPs (6). However, we see names that are integrated in sentence syntax or at the end of a tweet as purposefully added and thus count them as NPs (7). This holds for 179 of the 5K usernames in the **tw** corpus.

(6) @brycetache @TeresaMac2009 @BarackObama Thank you Obama!

(7) [@JoeNBC] just said twice [the @washingtonpost] deleted his unnamed quotes.

There are 205 instances of hashtags in **tw**, which we all consider as NPs. However, handling 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 (8)). 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 (8)), 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.

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

After detecting the NP structures, we differentiated the personal pronouns (PRPs) by the PoS categories in **ont** and **swbd**, and by pronoun types in gold coreference annotation in the **tw** corpus. We classify their ”person” feature (i.e. 1st, 2nd, 3rd) by string matching. Pronoun variants stemming from transcription (e.g.,”yo-”,”em”) or speaker’s choices such as contracted forms of nominal pronouns in **tw** (e.g.,”im”,”hes”,”youve”) are also considered as valid pronoun forms.

4.2 Frequency-based Features

As frequency-based metrics for the genre/medium comparison, we computed 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 Table 2 gives descriptive statistics for NP and PRP instances, and Figure 1 shows the distribution of large span NPs and PRPs. For confirming the statistical significance of differences in frequency data, we applied Pearson's χ^2 with post-hoc pairwise Fischer test where the correction method for multiple comparison is set to "holm". The tests confirmed that the differences between genres are statistically significant (p-value<0.05). The representative chart for the pronoun distribution in terms of grammatical person is shown in Figure 2. All the differences in personal pronoun distribution between genres except the *swbd-tc* pair are statistically significant (p-value<0.05).

	tw	**swbd**	**OntoNotes**				
Feature			**tc**	**bc**	**bn**	**nw**	**wb**
# of Personal Pronouns	3068	4812	5210	3158	1590	788	1659
# of 1st Person Pronouns	1066	1921	2130	1218	457	134	624
# of 2nd Person Pronouns	706	1055	1077	594	175	37	194
# of 3rd Person Pronouns	1292	1835	2002	1343	955	611	829
# of Noun Phrases[6] (Large Span)	6078	4323	4121	4955	5647	564	4733
# of Noun Phrases (Short Span)	7576	5537	4855	7281	8642	8607	7896
# of Personal PRPs in Parentheticals	37	670	530	161	40	17	19
Nominal Density (SS) (%)	26.6	25.9	25.2	26.1	25.6	23.5	23.9

Table 2: Descriptive frequency statistics per 40K tokens (normalized values)

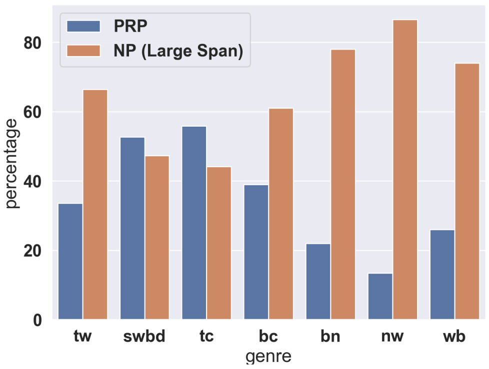


Figure 1: Distribution of Large Span NPs and PRPs

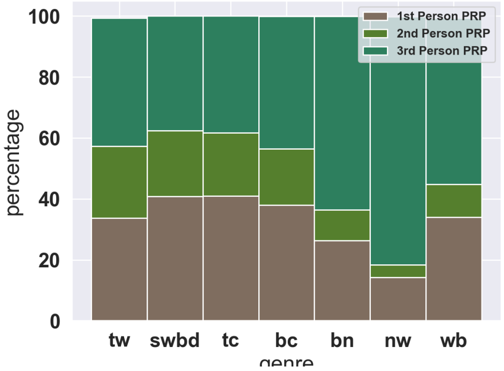


Figure 2: Distribution of Personal Pronouns

4.3 Heaviness-based Features

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 other 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. 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 count 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. It is

[6] Personal pronouns are not included in this count.

computed with the constituency parses in **ont** and **swbd**, while for **tw**, we run the Stanford parser on gold NP spans and use the automatically created parse trees for calculating NP-Height. We do not apply manual verification to the parse trees for this measure. We compute the heaviness metrics for both large and short span NPs, and we exclude personal pronouns in heaviness-based comparison of NPs.

Results The average NP-Length and NP-Height values across genres are shown in Table 3. The NP-height and NP-length data do not follow the normal distribution. Therefore, we applied a non-parametric statistical test (Kruskal-Wallis test) to assess the significance of differences among genres. For pairwise comparison of the genres, we applied Wilcoxon rank sum test. The tests indicate that differences in average NP-Length values are statistically significant for all the pairs except *tw-tc* for large span NPs and *bn-wb* pair for short span NPs (p-value<0.05). 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).

	tw	**swbd**	**OntoNotes**				
Feature			**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	3.39	4.64	4.67
NP-Height (Short Span)	3.11	3.06	3.06	3.07	3.09	3.13	3.09

Table 3: Average NP-Length and NP-Height across datasets

4.4 Distance-based Features

We measured the linear distance between anaphoric 3rd person pronouns and their antecedents in terms of tokens, clauses and noun phrases. We excluded 1st and 2nd person pronouns, and non-anaphoric intensifier self-forms of 3rd person pronouns (e.g. ”The prisoner **himself** can come to the points”) in distance-based computations.

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 (e.g. in example (9) there are 1699 tokens between two instances of ”they”, but an additional mention for that entity -*channel thirteen*- was mistakenly 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.

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

Token-based Distance 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 3.4, the hashtags and usernames that are not automatically inserted by the UI, 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 1). To see the impact of these tokens on TBD, we additionally measured the distance without considering discourse markers (TBD$'$).

Clause-based Distance 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 in the same way as done by Aktaş et al. (2019). We manually correct the identified clause boundaries for **tw** corpus due to low accuracy of automatically created parse trees. Automatically inserted @-usernames are not considered as part of the clauses in **tw**. 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”. We marked the first token of each identified clause and counted the number of marked tokens between two mentions to calculate the CBD. As shown in Table 1, parenthetical clauses (PRNs) are frequent in spoken genres. To see their impact on CBD, we additionally measured the distance without considering parentheticals (CBD′).

NP-Based Distance 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′).

Results Table 4 shows the average values for distance metrics. 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 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 statistical 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 the differences in CBD are statistically significant (p-value<0.05) for both settings. For NBD, except from the *bn-wb*, *nw-wb* pairs, all the 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.

	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: Average distance measures across datasets

5 Discussion

”Spoken” and ”written” are in one sense trivial, but from a linguistic perspective fairly problematic concepts. Following the notion of ”conceptual orality” by Koch and Oesterreicher (1985), we do not assume a binary distinction but a continuum from ”close” to ”distant” language, which only loosely corresponds to the two media. In OntoNotes, for example, the ”broadcast news” genre 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 spontaneous spoken genres swbd/tc and two written genres nw/wb are always located closely (if not adjacent) in the ranking of genres in terms of the average values of the features investigated, and the two groups are situated at the opposite ends.

Our frequency-based analysis shows that the relative frequencies of pronouns and NPs in the swbd/tc pair are close to each other, whereas NPs are much more dominant in nw/wb. The comparison of distance-based features indicates that the textual distances between anaphoric pronouns and their antecedents are longer in swbd/tc than nw/wb. And lastly, nw/wb genres contain heavier NPs than swbd/tc do. These inter-medium differences are statistically significant for all the measures except TBD, whereas for a number of cases, as demonstrated in Section 4, intra-medium differences were not confirmed statistically. The significance levels can differ when parentheticals and discourse markers are excluded in the computation of features, but these differences do not affect the statistical significance levels 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 close to the swbd/tc pair with respect to NP-Length, but to nw/wb pair with respect to NBD.

In addition to rankings derived from average-value orderings, we also used **hierarchical clustering** for grouping the genres based on the quantitative features. To obtain the clusters, we first normalized the data size, and used 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 applied the *hclust* method in R with default settings. Similar to the observations mentioned above, the swbd/tc and nw/wb pairs are always grouped together for all the metrics except NBD for swbd/tc and NP-Height for nw/wb. Again, for bc and tw the grouping differs from feature to feature. For instance, in Figure 3, bc is clustered in the same branch with nw/wb, whereas in Figure 4, it is grouped together with the swbd/tc pair.

The ranking-based and cluster-based groupings do not always overlap (as in the NBD 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: bn genre is usually grouped with the nw/wb pair, as it is expected due to the edited content of the genre. The only exception is in hierarchical clustering with respect to the frequencies of personal pronouns, where bn is clustered together with swbd/tc. We consider these findings (i.e. existence of more rigid clusters as well as the floating genres between the groups) as supportive evidence for perceiving spoken/written media as a continuum rather than discrete concepts, and it turns out that Twitter has a medium position in the continuum w.r.t. to the investigated features. The exact position differs from feature to feature.

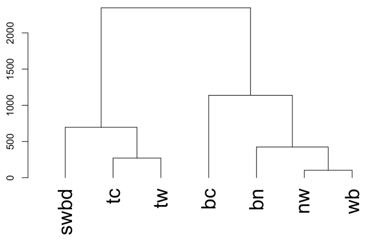


Figure 3: Genre clustering based on CBD

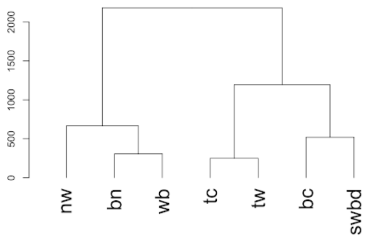


Figure 4: Genre clustering based on NP-Length (LS)

6 Conclusions and Future work

We presented an in-depth study of coreference strategies across genres that involve different production media (spoken, written). As a prerequisite, we harmonized the annotations (for syntax and coreference) 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 have been given in the previous section. Some of the features used in this paper such as pronoun and NP frequency, the distribution of personal pronouns, and the length of noun phrases are also considered in more general discussion of spoken and written language (Mair, 2006; Wasow and Arnold, 2011; Jonsson, 2016) and our results are compatible with these previous findings. Although we find the discussion about fine-grained characteristics of genre and medium (e.g. ”involvement” vs ”detachment” (Chafe, 1982)) interesting and important, considering the scope of this paper and size constraints, we leave it as a future work. In addition, we see our results as potentially fruitful for adapting automatic coreference resolution to genres and media, in the light of small amounts of training data, where knowledge about the differences in strategies could be utilized. Such experiments are a part of our future work, as is a psycholinguistic study (text production experiment) designed to validate the results on the spoken-written contrast from a perspective that is complementary to corpus analysis.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions, which have helped us significantly improve the paper. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 317633480 - SFB 1287, Project A03.

References

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, June. Association for Computational Linguistics.

Berfin Aktaş, Tatjana Scheffler, and Manfred Stede. 2019. Coreference in English OntoNotes: Properties and Genre Differences. In Kamil Ekštein, editor, *Text, Speech, and Dialogue*, pages 171–184, Cham. Springer International Publishing.

Berfin Aktaş, Veronika Solopova, Annalena Kohnert, and Manfred Stede. 2020. Adapting Coreference Resolution to Twitter Conversations. In *Findings of EMNLP*. Association for Computational Linguistics.

Marilisa Amoia, Kerstin Kunz, and Ekaterina Lapshinova-Koltunski. 2012. 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. European Language Resources Association (ELRA).

BBN Technologies, 2007. *Co-reference Guidelines for English OntoNotes Version 7.0.*

Douglas Biber, Edward Finegan, Stig Johansson, Susan Conrad, and Geoffrey Leech. 1999. *Longman Grammar of Spoken and Written English*. Longman.

Douglas Biber. 1992. 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.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 80–87, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. 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.

Wallace L. Chafe. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen, editor, *Spoken and Written Language: Exploring Orality and Literacy*, pages 35–53. Norwood, NJ: Ablex.

Miriam Eckert and Michael Strube. 2000. Dialogue Acts, Synchronizing Units, and Anaphora Resolution. *Journal of Semantics*, 17(1):51–89.

S. Engell. 2016. *Coreference in English and German: A Theoretical Framework and Its Application in a Study of Court Decisions*. Logos Verlag Berlin.

Barbara A. Fox. 1987. *Discourse structure and anaphora : written and conversational English / Barbara A. Fox.* Cambridge University Press Cambridge [Cambridgeshire] ; New York.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, volume 1, pages 517–520.

Christian Hardmeier, Luca Bevacqua, Sharid Loáiciga, and Hannah Rohde. 2018. 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. Association for Computational Linguistics.

Ewa Jonsson. 2016. *Conversational Writing: A Multidimensional Study of Synchronous and Supersynchronous Computer-mediated Communication*. English Corpus Linguistics. Peter Lang, Frankfurt am Main.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe - Sprache der Distanz. Müdlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(1985):15 – 43.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. 2015. 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.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, and José Manuel Martínez. 2016. 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, San Diego, California, June. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski. 2015. Exploration of inter- and intralingual variation of discourse phenomena. In *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT@EMNLP)*, pages 158–167, Lisbon, Portugal, September. Association for Computational Linguistics.

Christian Mair. 2006. *Twentieth-Century English: History, Variation and Standardization*. Studies in English Language. Cambridge University Press.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. 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. Association for Computational Linguistics.

Stella Neumann and Jennifer Fest. 2016. 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.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica Macbride, and Linnea Micciulla. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. *International Conference on Semantic Computing*, 0:446–453, 09.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152. Association for Computational Linguistics.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop*, pages 57–62, Berlin, August. Association for Computational Linguistics.

Catherine Schnedecker. 2018. 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.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Sapporo, Japan, July. Association for Computational Linguistics.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: An Overview. In Abeillé A., editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*. Springer, Dordrecht.

Olga Uryupina and Massimo Poesio. 2012. 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. European Language Resources Association (ELRA).

Thomas Wasow and Jennifer Arnold, 2011. *Post-verbal constituent ordering in English*, pages 119–154. De Gruyter Mouton, 01.

Thomas Wasow. 1997. Remarks on grammatical weight. *Language Variation and Change*, 9:81 – 105, 03.

Amir Zeldes. 2018. 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. Association for Computational Linguistics.