Coreference in Spoken vs. Written Texts: a Corpus-based Analysis

Marilisa Amoia, Kerstin Kunz, Ekaterina Lapshinova-Koltunski

Department of Applied Linguistics, Saarland University
{m.amoia,k.kunz,e.lapshinova}@mx.uni-saarland.de

Abstract

This paper describes an empirical study of coreference in spoken vs. written text. We focus on the comparison of two particular text types, interviews and popular science texts, as instances of spoken and written texts since they display quite different discourse structures. We believe in fact, that the correlation of difficulties in coreference resolution and varying discourse structures requires a deeper analysis that accounts for the diversity of coreference strategies or their sub-phenomena as indicators of text type or genre. In this work, we therefore aim at defining specific parameters that classify differences in genres of spoken and written texts such as the preferred segmentation strategy, the maximal allowed distance in or the length and size of coreference chains as well as the correlation of structural and syntactic features of coreferring expressions. We argue that a characterization of such genre dependent parameters might improve the performance of current state-of-art coreference resolution technology.

Keywords: Coreference Resolution, Corpus-based analysis, Spoken Language

1. Introduction

Coreference resolution, i.e. the identification of referring expressions that point to the same extralinguistic referent in a natural language discourse still remains one of the most demanding tasks in NLP. This was clearly highlighted by various coreference resolution challenges such as MUC-6 (1995), MUC-7 (1997) or ACE NIST (2004) (see (Doddington et al., 2004)) and more recently the CoNNL 2011 (see (Pradhan et al., 2011)) shared task. Interestingly, the baseline for state-of-art systems in the early MUC challenges was about 60-70% and it increased in the ACE 2004 to about 85%. The performance [1] significantly decreased in the CoNNL 2011 challenge, where the best system achieved 57.79%. An explanation for this decrease is not only that more complex aspects of coreference were included in the task, e.g. the integration of non-nominal reference. We presume that the extension of the dataset to spoken language genres such as broadcast news and conversations is one major factor for low performance rates as most coreference resolution systems have been developed for processing written text input.

[1] These values of system performance are based on different metrics: MUC applies the link measure (Vilain et al., 1995), ACE a value-based metric called the ACE-value (Doddington et al., 2004) and ConLL the unweighted mean of the MUC, B_CUBED (Bagga and Baldwin, 1998) and $CEAF_e$ (Luo, 2005) measures. However, we think that the differences are still indicative. Consider, for instance, the performance scores (59.57 and 57.79) of the CoreNLP system for the metrics MUC and ConLL respectively as reported in (Pradhan et al., 2011).

It is not a new issue that difficulties in coreference resolution greatly increase when spoken language is involved. There are some interesting works that propose algorithms for coreference resolution in spoken language. For instance, (Eckert and Strube, 2000) analyze the frequency distribution of personal and demonstrative anaphora in spontaneous speech dialogues and use dialogue act segmentation to improve pronoun resolution. (Tetreault and Allen, 2004) apply an algorithm based on the Question Under Discussion technique for coreference resolution of pronouns in task-oriented dialogues and (Stent and Bangalore, 2010) propose a statistical coreference system for the same domain using a stack model for representing the intensional discourse state.

We however believe that the correlation between difficulties in coreference resolution and varying discourse structures requires a deeper analysis that accounts for the diversity of coreference strategies or their sub-phenomena as indicators of text type or genre. In this work, we therefore aim at defining specific parameters that classify differences in genres of spoken and written texts such as the preferred segmentation strategy (e.g. paragraph, turn, speech act, etc.), the maximal allowed distance in or the length and size of coreference chains as well as the correlation of structural and syntactic features of coreferring expressions. We argue that a characterization of such genre dependent parameters might improve the performance of current state-of-art coreference resolution technology.

In this paper, we present an in-depth empirical analysis of the coreference strategies identified by the Stanford CoreNLP system (Lee et al., 2011). We focus

on the comparison of two particular text types, interviews and popular science texts, as instances of spoken and written texts since they display quite different discourse structures. We propose some metrics that might yield a more differentiated classification of the coreference properties in written vs. spoken language genres than has been offered by previous studies. Furthermore, we show that the difference in performance of a state-of art coreference extraction system run on both text genres can be explained on the basis of these differing structural properties.

2. Defining the Task: Coreference Resolution

Coreference involves a textual relation that is created between linguistic expressions. This textual relation evokes a conceptual relation of identity between discourse referents. A coreference relation links at least two linguistic expressions: an antecedent, i.e. a linguistic element introducing a new discourse referent, and an anaphor (or cataphor, in the case of forward reference) pointing to the same referent again. Thus, a **coreferring expression** may either be an antecedent or an anaphor, pointing to a referent that is also mentioned at another point in the same discourse and a **coreference chain** is the set of all the coreferring expressions which refer to the same antecedent. Coreference resolution is all about tracking the complete set of coreference chains in a text.

Most works on coreference focus on the analysis of anaphors as their specific linguistic features are considered to trigger the coreference relation to the antecedent. In our study we provide a fine-grained classification of properties of all coreferring expressions (i.e. antecedents and anaphors) along with a detailed characterization of coreference chains.

More precisely, we focus on the following aspects of coreference:

- Lexical coreference: includes all cases of nominal coreference to the same entity through the lexical means such as named entity (NE), head nouns in singular and plural employed as repetitions, synonyms, hyponyms etc. e.g. *fork, spoon* and *cutlery* in (1). Note that coreferring lexical phrases may be introduced by other means of reference, such as articles and demonstrative determiners and additionally contain modifying elements, such as adjectives (see 2) or relative clauses.

  (1) a. *Ann put a* **fork** *and a* **spoon** *on the table.*
  b. *But she forgot the* **cutlery** *for her mother.*

  (2) a. **Ann** *is the daughter of Mary.*
  b. *I met* **this beautiful girl** *in the bus yesterday.*

- Pronominal coreference: includes reference by personal pronouns, demonstrative pronouns, possessive pronouns and their morphological variants, e.g. *Mark and John* in (3a), *they* in (3b) and *their* (3c).

  (3) a. *Yesterday* **Mark and John** *entered the bank around the corner.*
  b. **They** *were masked and were carrying guns.*
  c. **Their** *guns were loaded.*

3. Evaluation Framework

For our empirical analysis of coreference strategies in different text genres we use a corpus of written and spoken texts (see 3.1.). The corpus is annotated with the deterministic coreference resolution system provided by Stanford CoreNLP (Lee et al., 2011) (see 3.2.). To describe the coreference strategies observed in the corpus, we define a number of metrics described in 3.3.

3.1. Corpus Resources

For the analysis we use a corpus of English including two text genres:

- POPSCI: includes 11 texts on popular science journals composed by experts for an educated lay audience. The written texts are prepared and monologic, hence, there is no direct contact and interaction between speech participants. The topics treated are content-oriented, dealing with scientific phenomena.

- INTERVIEW: includes 11 manually transcribed oral interviews. The texts were gathered from the Backbone Corpus (Kohn et al., 2009). The selected interviews are dialogic, all speech participants are physically present in the speech situation, thus, direct contact and verbal interaction is given between speech participants. There is an interviewer, who poses prepared questions about everyday life. Interposed questions are spontaneous. The interviewees are native speakers of American English. Their answers are not prepared but spontaneous. The interviewees have a larger share in the overall verbal interaction than the interviewers. The topics dealt with are speaker-oriented and center around private and professional life of the interviewees.

The whole corpus contains about 140000 tokens and each text file includes about 6000 tokens. The spoken language texts are manually preprocessed, i.e. the errors have been rewritten and the texts have been normalized and segmented into sentences. Furthermore, spoken language features such as fillers, corrections and repetitions are encoded by XML-tags. This specific design allows for an automatic analysis with NLP techniques but still captures main characteristics of spoken language.

3.2. Coreference Annotation

The corpus is annotated with the deterministic coreference resolution system provided by Stanford CoreNLP (Lee et al., 2011). This system achieved the best performance in the CoNLL 2011 unrestricted coreference resolution task. CoreNLP implements an incremental strategy: In order to identify coreferring expressions, first the set of possible mentions (i.e. coreferring items) is identified. Then, in the next successive steps, the system tries to refine this initial set by pruning mentions which are no more consistent with the model. This is achieved by applying different sieves in sequence. Each sieve applies a deterministic model of coreference, e.g. string matching, mention similarity or pronoun resolution. The sieves are applied in order from the highest to lower precision. At each step the system uses as input the results of the preceeding sieves.
The system deals with nominal and pronominal coreference but does not implement discourse-deictic reference, i.e. it does not account for sentential or VP antecedents. Moreover, from the initial set of noun phrases, named entities and pronouns in the set of possible mentions, those of the type listed below (among others) are pruned and are thus not accounted for by the current version of CoreNLP:

(i) adjectival forms of nations,

(ii) pleonastic-it annotations recognized by means of recurrent/general patterns, e.g. *it is possible*, etc.

(iii) expressions including numbers such as cardinals, percents, etc.

3.3. Metrics for Comparison

In order to characterize structural differences between texts in written and spoken genres we used the following metrics:

- **Precision**, percentage of rightly annotated out of all annotated coreferring expressions.
- **T-length**, average token length of coreferring expressions.
- **Chain Size**, i.e. number of coreferring expressions in one chain.
- **S-distance**, average number of sentences separating coreferring expressions in the same chain.
- **Parallelism**, number of coreferring anaphors in one chain that exhibit the same or similar syntactic features as the antecedent (e.g. sentence-initial subject). Parallelism captures the similarity between the syntactic context of the anaphors and their antecedent, e.g.

  (4) a. Ann eats many apples.
  b. She likes them.

- **Grammatical role preference** of different types of coreferring expressions, e.g. coreferring pronouns that are realized as subjects in the sentence in which they occur or coreferring lexical phrases that appear as syntactic objects. The grammatical roles investigated are subject and object. In addition, coreferring expressions may appear in lower ranks of the sentence and be embedded in phrases. They are then classified as modifiers.
- **Typology** of coreferring expressions. We distinguish lexical coreference (named entities and full lexical noun phrases) and different types of pronominal reference (personal, possessive and demonstrative pronouns).
- **Morphological Features** of coreferring expressions (singular vs. plural, 1st, 2nd or 3rd person).

4. Results

In this section, we outline the results of our preliminary analysis of the collected data. The values observed for the metrics described in section 3.3. for the written as well as for the spoken corpus are displayed in Table 1 to Table 5. The differences in the values are all statistically significant with $p < 0.05$ (Student's t-Test). We start by a discussion of some general features as shown in Table 1.
In order to determine the precision of the coreference annotation, we have manually corrected 4 out of the 22 texts (2 texts from each subcorpus) that were tagged by the CoreNLP system. The estimated precision[2] of the recognized referring expressions in spoken (51%) vs. written texts (64%) confirms the results of previous work in that automatic coreference resolution systems differ in their performance to process spoken

[2] As evaluation metric, we use the link-based MUC metric (Vilain et al., 1995).

	POPSCI	INTERVIEW
Precision		
Average	64%	51%
Tokens per Sentence		
Average	25.75	20.73
Chain Size		
Average	3.60	4.45
T-Length		
Average	3.42	2.58

Table 1: Comparison of Surface Features.

vs. written texts, and in that the tools perform better on written than on spoken language. We presume that these differences in performance can be correlated with the different strategies employed to realize coreference in written vs. spoken texts.

Indeed, we found evidence in our data that certain factors might be crucial for modeling the structural diversity of different text genres and thus might be used as a parameter for enhancing automatic reference resolution. For instance, a comparison of the values we gathered for surface features, such as those summarized in Table 1, shows that written text sentences are typically longer than those uttered in spoken language (25.75 vs. 20.73). The same trend can be observed for the T-length parameter, written text coreferring expressions are generally longer than the ones used in spoken language (3.42 vs. 2.58).

This tendency is further confirmed by the observation that, in written texts, lexical coreference (e.g. NE, repetitions) is preferred over pronominal coreference (cf. Table 2). On the contrary, pronominal reference clearly is the most frequent strategy in spoken genres. Note that Table 2 only reports the percentage of lexical and pronominal coreference. Their sum is not 100. The rest, i.e. 7.9% for the written and 10.1% for the spoken texts are coreferring expressions of the following types: adjective, adverbs, cardinals, etc. We consider these annotations as errors, as the CoreNLP system is not supposed to annotate other coreference types than nominal and pronominal.

Spoken language constrains referential elements to span shorter (6.5 sentences) textual distance (s-distance between anaphor and antecedent) if compared with written text (8.8 sentences in average) (cf. Table 3). However, lexical coreference seems to allow segments of equal length in both text genres and generally might involve very long spans of text, circa 12 sentences, as compared to 2.5 sentences observed for pronominal reference. These effects might probably be explained by constraints on short term memory capacity, which manifest to a greater degree in spoken than in written language. As for the chain size displayed in Table 3, we find that coreference chains in the spoken texts contain more coreferring expressions than the written texts. This is strongly related to speaker orientation, which translates in a high number of first person pronouns, most of which are contained in one and the same coreference chain.

	POPSCI	INTERVIEW
Lexical Coreference		
Named Entity	14.1%	7.1%
Noun Phrase	48.9%	25.0%
TOT	63.0%	32.1%
Pronominal Coreference		
Personal Pron	18.4%	51.1%
Possessive Pron	7.4%	5.2%
Demonstrative Pron	3.3%	1.5%
TOT	29.1%	57.8%

Table 2: Distribution and Typology of CorefType.

	POPSCI	INTERVIEW
S-Distance		
Average	8.8	6.5
Lex	11.3	12.6
Pron	2.5	2.6
This	1.5	2.3

Table 3: S-distance and CorefType

	POPSCI	INTERVIEW
Lexical Coreference		
Subj	41.13%	31.75%
Obj	24.46%	53.9%
Mod	34.41%	14.29%
Pronominal Coreference		
Subj	86.08%	97.19%
Obj	11.39%	2.81%
Mod	3.0%	0.0%

Table 4: Reference Type and Grammatical Case.

Table 4 shows that coreferring pronouns are preferably used in subject position in both text genres and less often in other syntactic roles. However variation is greater in the written texts. In the spoken texts, almost all coreferring pronouns are subjects and none do appear in modifying phrases. This is due to the speaker-centered presentation of utterances which typ-

ically manifests in heavy usage of first and second person pronouns (see below). Lexical coreference strategies exhibit a wider range of grammatical applications in the spoken texts than pronouns as coreferring lexical phrases occur as subjects, objects or as a modifiers. Quite interestingly, the highest distribution in INTERVIEW is found for objects, followed by subjects. Yet, this can be explained by the fact that old and highly salient information in INTERVIEW is typically represented by pronouns in sentence-initial subject position (see below) and that newer and less salient information realized by lexical phrases is placed in post-verbal object position. Note that many of the lexical noun phrases are made up of general nouns such as *people, children, schools*. They thus lack semantic specificity and yield a higher referential ambiguity, as compared to POPSCI. Distributions in POPSCI differ largely from INTERVIEW as most coreferring lexical noun phrases appear as subjects, and a fairly high distribution is traced for modifiers. The latter is due to the fact that, in POPSCI, noun phrases generally have a high structural complexity, hence coreferring noun phrases often function as modifiers of superordinated NPs. In addition, coreferring lexical noun phrases in POPSCI exhibit a high ontological specificity. They are often terminological expressions pointing to domain-specific referents. The lower specificity and higher ambiguity/vagueness tracked in lexical phrases in INTERVIEW can be identified as one reason for lower precision in comparison to POPSCI. This particularly holds in cases where orthographical repetition of lexemes is involved: IN POPSCI, repetitions of terminological expressions often evoke coreference (see example 5a and 5b below).

(5) a. *The most sophisticated robotic spacecraft ever built,* **the Cassini orbiter** *and* **the attached Huygens probe***, were poised atop the launch vehicle, ...*

b. *Whereas Galileo released a probe to investigate Jupiter's atmosphere,* **the Cassini orbiter** *will send* **the Huygens probe** *to Titan, not Saturn.*

As example 6a to 6c illustrates, repetitions of general nouns in INTERVIEW are often wrongly assigned to the same coreference chain.

(6) a. **the children** *range from the age of about 3 which is nursery, going up to about 10 or 11 in Year 6...*

b. *And* **the children** *and myself are both noticing that, so...*

c. *It's a nice subject and I think it's a very essential subject for* **the children** *nowadays, definitely.*

The spoken texts also display a higher number of parallel syntactic constructions in coreference chains if compared to written texts (16% of all constructions in the spoken vs. 6% in the written). This is mainly due to, first, a higher distribution of canonical sentence structures and, second, a less marked positioning of given/salient and new/less salient information in general. As a result, we found a high number of pronominal subjects occurring in sentence-initial position. Hence, in spoken texts, the most frequent parallel syntactic constructions have the structure NP1 VP NP2 or PP, where NP1 is a subject and NP2 is an object (see e.g. example 7a and 7b), and PP an adverbial complement .

(7) a. **I** *did my placement at another school in Sutton which was lovely, a very nice school, called Westbourne actually.*

b. *And* **I** *did another placement at a school in Leatherhead, so I did 2 placements in my 1 year.*

This type of construction is less frequent in the written discourse, where we observe more variation in the organization of syntactic constituents, and hence, less parallel constructions. This may be due to the fact that sentences generally contain more new and less highly salient information as compared to INTERVIEW. For instance, the preverbal position is often filled with more than one syntactic constituent, resulting in a PP NP VP construction or in clause NP VP (see example 8a-8c).

(8) a. *However, at that time* **interferons** *had never been properly tested in specially designed clinical trials on large numbers of patients, because the drugs were available only in minute quantities.*

b. *Although many elegant studies were done and* **the drug** *looked promising it was never possible to treat enough people with sufficient interferon to be sure of the results.*

c. *However, these scanty data and the public interest in* **interferon** *led to a great commercial impetus to produce enough interferon to adequately test on people.*

Pronouns: Morphological Types		
	POPSCI	**INTERVIEW**
1st per sg	5.8 %	26.9%
1st per pl	15.5 %	17.8%
3rd per sg	45.1 %	37.7%
3rd per pl	33.1 %	16.9%
2nd per	0,5 %	0.8%

Table 5: Morphology of personal and possessive pronouns.

Other frequent constructions are of the type NP VP1 VP2, where VP2 is expressed by an infinitive, or of the type NP VP PP. Note again that coreferring expressions in POPSCI frequently occur as modifiers because of heavy phrase embedding (consider 7a-c). The comparison of structural diversity across spoken and written texts shows that the diversity in the spoken texts is not as rich as in the written ones. This can be partially explained by less variation in the type of coreferential device employed.

Additionally, the two described tables interdepend with the features indicated in Table 5: the distinct morphological features of coreferring personal and possessive pronouns in spoken vs. written texts. There are two main contrasts to be highlighted here: First, we trace a much higher number of first person pronouns (most notably in singular but also in plural) in INTERVIEW than in POPSCI, which reflects the speaker centered textual function of this text type (see again example 6a/b). Note that the first person pronouns are often realized as sentence-initial subjects. In case of speaker turns they belong to different coreference chains, which is one cause for a reduced precision rate. The difference in distribution of first person pronouns between the two genres goes along with the second contrast we observe: third person pronouns are less frequent in INTERVIEW than in POPSCI. One interesting observation not included in Table 4 is that the distribution of the neuter pronoun *it* is much higher in the INTERVIEW corpus (90.8 % of all third person pronouns in singular and 34% of all pronouns) than in POPSCI (88 % of all third person pronouns in singular and 40% of all pronouns). Our findings hereby confirm those described by (Eckert and Strube, 2000) for spoken texts. The grammatical function of *it* varies from a dummy subject and reference to (non)-nominal antecedents to vague reference to an antecedent that is not clearly defined. Consider 8, where *it* in 8b and 8d are coreferring with *Reigate*, but not 8c.

(9) a. *I live in a town called called* **Reigate**.

b. **It***'s between London and the countryside which is quite nice.*

c. **It** *takes us about 25 minutes to get to London on the train.*

d. **It***'s I say* **it***'s a town,* **it***'s more of a village.*

Vague reference also is quite frequent for usage of the third person plural form:

(10) a. *So you can go and explore these places, a bit like a museum in a way. How does it work language-wise? If you, say, if you go to Rome or to Spain or whatever, is it all in English or is it how do* **they** *do this?*

b. *That's a good question. In some areas you might* **they** *might have just one language spoken,...*

Our observations outlined above are assumed to be one reason for lower precision in INTERVIEW as compared to POPSCI.

5. Conclusions

In this paper we have presented a preliminary analysis of the different coreference mechanisms that prevail in written vs. spoken texts. We have classified the coreference phenomena encoded with a state-of-art coreference resolution system according to their morphological, syntactical and relational features. This was done in two particular genres, interviews and popular science texts, to obtain initial parameters for an algorithm that allows analyzing written and spoken texts alike. Our findings reveal that different parameters have to be defined for the two genres as they exhibit quite distinct features: written texts call for a deeper analysis of embedded structures along with an ontological classification of lexical coreferring expressions, whereas spoken texts require a fine-grained differentiation of structural and functional types of pronouns in combination with an annotation of speaker turns. Furthermore a semantic analysis of the context in which

the coreference items occur might improve coreference resolution for lexical coreference with general nouns. We intend to extend the analysis to other spoken genres (e.g. presentations and conversations) and incorporate diversity among spoken language genres in order to identify parameters for a coreference algorithm with which systematic differences between written and spoken texts can be captured.

6. References

A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LRECÕ98)*, page 563Ð566.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) program tasks, data, and evaluation. In *NIST*.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Kurt Kohn, Petra Hoffstaedter, and Johannes Widmann. 2009. Backbone - pedagogic corpora for content and language integrated learning. In *Eurocall-2009*, Valencia, Spain, September 9-12.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. In *CoNLL-2011 Shared Task*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *HLT/EMNLP*.

MUC-6. 1995. Coreference task definition. In *the Sixth Message Understanding Conference (MUC-6)*.

MUC-7. 1997. Coreference task definition. In *the Seventh Message Understanding Conference (MUC-7)*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unre- stricted coreference in ontonotes. In *the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*.

Amanda J. Stent and Srinivas Bangalore. 2010. Interaction between dialog structure and coreference resolution. In *IEEE SLT 2010*, Berkeley, USA.

Joel Tetreault and James Allen. 2004. Dialogue structure and pronoun resolution. In *DAARC'04*, Azores, September 23-24.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52.