Using prosodic annotations to improve coreference resolution of spoken text

Ina Rösiger and Arndt Riester
Institute for Natural Language Processing
University of Stuttgart, Germany
Pfaffenwaldring 5b, 70569 Stuttgart
roesigia|arndt@ims.uni-stuttgart.de

Abstract

This paper is the first to examine the effect of prosodic features on coreference resolution in spoken discourse. We test features from different prosodic levels and investigate which strategies can be applied. Our results on the basis of manual prosodic labelling show that the presence of an accent is a helpful feature in a machine-learning setting. Including prosodic boundaries and determining whether the accent is the nuclear accent further increases results.

1 Introduction

Noun phrase coreference resolution is the task of determining which noun phrases (NPs) in a text or dialogue refer to the same discourse entities (Ng, 2010). Coreference resolution has been extensively addressed in NLP research, e.g. in the CoNLL shared task 2012 (Pradhan et al., 2012) or in the SemEval shared task 2010 (Recasens et al., 2010). Amoia et al. (2012) have shown that there are differences between written and spoken text wrt coreference resolution and that the performance typically drops when systems that have been developed for written text are applied on spoken text. There has been considerable work on coreference resolution in written text, but comparatively little work on spoken text, with a few exceptions of systems for pronoun resolution in transcripts of spoken text e.g. Strube and Müller (2003), Tetreault and Allen (2004). However, so far, prosodic information has not been taken into account. The interaction between prosodic prominence and coreference has been investigated in several experimental and theoretical analyses (Terken and Hirschberg, 1994; Schwarzschild, 1999; Cruttenden, 2006); for German (Baumann and Riester, 2013; Baumann and Roth, 2014; Baumann et al., 2015).

There is a tendency for coreferent items, i.e. entities that have already been introduced into the discourse, to be deaccented, as the speaker assumes the entity to be salient in the listener's discourse model. We can exploit this by including prominence features in the coreference resolver.

Our prosodic features mainly aim at definite descriptions, where it is difficult for the resolver to decide whether the potential anaphor is actually anaphoric or not. In these cases, accentuation is an important means to distinguish between given entities (often deaccented) and other categories (i.e. bridging anaphors, see below) that are typically accented, particularly for entities whose heads have a different lexeme than their potential antecedent. Pronouns are not the case of interest here, as they are (almost) always anaphoric. To make the intuitions clearer, Example (1), taken from Umbach (2002), shows the difference prominence can make:

(1) John has <u>an old cottage</u>.[1]
  a. Last year he reconstructed the SHED.
  b. Last year he reconSTRUCted **the shed**.

Due to the pitch accent on *shed* in (1a), it is quite obvious that *the shed* and *the cottage* refer to different entities; they exemplify a bridging relation, where the shed is a part of the cottage. In (1b), however, *the shed* is deaccented, which has the effect that *the shed* and *the cottage* corefer.

We present a pilot study on German spoken text that uses manual prominence marking to show the principled usefulness of prosodic features for coreference resolution. In the long run and for application-based settings, of course, we do not want to rely on manual annotations. This work is investigating the potential of prominence information and is meant to motivate the use of automatic

[1] Anaphors are typed in boldface, their antecedents are underlined. Accented syllables are capitalised.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 83–88,
Beijing, China, July 26-31, 2015.

prosodic features. Our study deals with German data, but the prosodic properties are comparable to other West Germanic languages, like English or Dutch. To the best of our knowledge, this is the first work on coreference resolution in spoken text that tests the theoretical claims regarding the interaction between coreference and prominence in a general, state-of-the-art coreference resolver, and shows that prosodic features improve coreference resolution.

2 Prosodic features for coreference resolution

The prosodic information used for the purpose of our research results from manual annotations that follow the GToBI(S) guidelines by Mayer (1995), which stand in the tradition of autosegmental-metrical phonology, cf. Pierrehumbert (1980), Gussenhoven (1984), Féry (1993), Ladd (2008), Beckman et al. (2005). We mainly make use of *pitch accents* and *prosodic phrasing*. The annotations distinguish *intonation phrases*, terminated by a major boundary (%), and *intermediate phrases*, closed by a minor boundary (-), as shown in Examples (2) and (3).

The available pitch accent and boundary annotations allow us to automatically derive a secondary layer of prosodic information which represents a mapping of the pitch accents onto a prominence scale in which the nuclear (i.e. final) accents of an intonation phrase *(n2)* rank as the most prominent, followed by the nuclear accents of intermediate phrases *(n1)* and prenuclear (i.e. non-final) accents which are perceptually the least prominent. To put it simply, the nuclear accent is the most prominent accent in a prosodic phrase while prenuclear accents are less prominent.

While we expect the difference between the presence or absence of pitch accents to influence the classification of short NPs like in Example (1), we do not expect complex NPs to be fully deaccented. For complex NPs, we nevertheless hope that the prosodic structure of coreferential NPs will turn out to significantly differ from the structure of discourse-new NPs such as to yield a measurable effect. Examples (2) and (3) show the prosodic realisation of two expressions with different information status. In Example (2), the complex NP *the text about the aims and future of the EU* refers back to *the Berlin Declaration*, whereas in Example (3), the complex NP *assault with lethal consequences and reckless homicide* is not anaphoric. The share of prenuclear accents is higher in the anaphoric case, which indicates lower overall prominence. The features described in Section 2.1 only take into account the absence or type of the pitch accent; those in Section 2.2 additionally employ prosodic phrasing. To get a better picture of the effect of these features, we implement, for each feature, one version for all noun phrases and a second version only for short noun phrases (<=4 words).

2.1 Prosodic features ignorant of phrase boundaries

Pitch accent type corresponds to the following pitch accent types that are present in the GToBI(S) based annotations.

Fall	H*L
Rise	L*H
Downstep fall	!H*L
High target	H*
Low target	L*
Early peak	HH*L
Late peak	L*HL

For complex NPs, the crucial label is the last label in the mention. For short NPs, this usually matches the label on the syntactic head.

Pitch accent presence focuses on the presence of a pitch accent, disregarding its type. If one accent is present in the markable, the boolean feature gets assigned the value *true*, and *false* otherwise.

2.2 Prosodic features including phrase boundary information

The following set of features takes into account the degree of prominence of pitch accents as presented at the beginning of Section 2, which at the same time encodes information about prosodic phrasing.

Nuclear accent type looks at the different degrees of accent prominence. The markable gets assigned the type *n2*, *n1*, *pn* if the last accent in the phrase matches one of the types (and *none* if it is deaccented).

Nuclear accent presence is a Boolean feature comparable to pitch accent presence. It gets assigned the value *true* if there is some kind of accent present in the markable. To be able to judge the helpfulness of the distinction between the categories that are introduced above, we experiment with two different versions:

(2) Anaphoric complex NP (DIRNDL sentences 9/10):

9: Im Mittelpunkt steht eine von der Ratspräsidentin, Bundeskanzlerin Merkel, vorbereitete “Berliner Erklärung”.

10:	Die Präsidenten [. . .] wollen	**[den**	**TEXT**	**über die**	**ZIEle**	**und**	**ZUkunft**	**der**	**EU]**	unterzeichnen.
	the presidents [. . .] want	[the	text	about the	aims	and	future	the	EU]	sign
		((	L*H		L*H-)	(	H*L		H*L	H*L -)%)
			pn		n1		pn		pn	

Central is the ’Berlin Declaration’ that was prepared by the president of the Council of the EU, Chancellor Merkel.
*The presidents want to sign [**the text about the aims and future of the EU.**]*

(3) Non-anaphoric complex NP (DIRNDL sentences 2527/2528):

2527: Der Prozess um den Tod eines Asylbewerbers aus Sierra Leone in Polizeigewahrsam ist [. . .] eröffnet worden.

2528:	[Wegen	KÖRperverletzung	mit	TOdesfolge		und	fahrlässiger	TÖtung]	MÜSsen	. . .
	[Due	assault	with	lethal consequence,		and	reckless	homicide]	must	
	((	H*L		L*H	-)	(		H*L -)%)		
		pn		n1				n2		

The trial about the death of an asylum seeker from Sierra Leone during police custody has started.
Charges include [assault with lethal consequence, and reckless homicide], . . .

1. Only *n2* accents get assigned *true*
2. *n2* and *n1* accents get assigned *true*

Note that a version where all accents get assigned *true*, i.e. *pn* and *n1* and *n2*, is not included as this equals the feature *Pitch accent presence*.

Nuclear bag of accents treats accents like a bag-of-words approach treats words: if one accent type is present once (or multiple times), the accent type is considered present. This means we get a number of different combinations ($2^3 = 8$ in total) of accent types that are present in the markable, e.g. *pn* and *n1* but no *n2* for Example (2), and *pn, n1* and *n2* for Example (3).

Nuclear: first and last includes linear information while avoiding an explosion of combinations. It only looks at the (degree of the) first pitch accent present in the markable and combines it with the last accent.

3 Experimental setup

We perform our experiments using the IMS HotCoref system (Björkelund and Kuhn, 2014), a state-of-the-art coreference resolution system for English. As German is not a language that is featured in the standard resolver, we first had to adapt it. These adaptations include gender and number agreement, lemma-based (sub)string match and a feature that addresses German compounds, to name only a few.[2]

For our experiments on prosodic features, we use the DIRNDL corpus[3] (ca. 50.000 tokens, 3221 sentences), a radio news corpus annotated with both manual coreference and manual prosody labels (Eckart et al., 2012; Björkelund et al., 2014)[4]. We adopt the official train, test and development split. We decided to remove abstract anaphors (e.g. anaphors that refer to events or facts), which are not resolved by the system. In all experiments, we only use predicted annotations and no gold mention boundary (GB) information as we aim at real end-to-end coreference resolution. On DIRNDL, our system achieves a CoNLL score of 47.93, which will serve as a baseline in our experiments. To put the baseline in context, we also report performance on the German reference corpus TüBa-D/Z[5] (Naumann, 2006), which consists

[2] To download the German coreference system, visit: www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/HOTCorefDe.html

[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/dirndl.html

[4] In this work, we have focused on improvements within the clearly defined field of coreference resolution, using prosodic features. As one of the reviewers pointed out, the DIRNDL corpus additionally features manual two-level information status annotations according to the *RefLex* scheme (Baumann and Riester, 2012), which additionally distinguishes bridging anaphors, deictic expressions, and more. Recent work on smaller datasets of read text has shown that there is a meaningful correspondence between information status classes and degrees of prosodic prominence, with regard to both pitch accent type and position (Baumann and Riester, 2013; Baumann et al., 2015). Moreover, information status classification has been identified as a task closely related to coreference resolution (Cahill and Riester, 2012; Rahman and Ng, 2012). Integrating these approaches is a promising, though rather complex task, which we reserve for future work. It might, furthermore, require more detailed prosodic analyses than are currently available in DIRNDL.

[5] http://www.sfs.uni-tuebingen.de/de/ascl/ressourcen/corpora/tueba-dz.html

System	CoNLL (+singl.)	CoNLL (-singl.)
IMS HotCoref DE (open)	60.35	48.61
CorZu (open)	60.27	45.82
BART (open)	57.72	39.07
SUCRE (closed)	51.23	36,32
TANL-1 (closed)	38.48	14.17

Table 1: SemEval Shared Task 2010 post-task evaluation for track *regular* (on TüBa 8), including and excluding singletons

System	CoNLL
IMS HOTCoref DE (no GB matching)	51.61
CorZu (no GB matching)	53.07

Table 2: IMS HotCoref performance on TüBa 9 (no singletons), using regular preprocessing

of newspaper text. In a post-task SemEval 2010 evaluation[6] our system achieves a CoNLL score of 60.35 in the *open, regular* track[7] (cf. Table 1). On the newest dataset available (TüBa-D/Z v9), our resolver currently achieves a CoNLL score of 51.61.[8] Table 2 compares the performance of our system against CorZu (Klenner and Tuggener, 2011; Tuggener and Klenner, 2014), a rule-based state-of-the-art system for German[9](on the newest TüBa dataset).

4 Experiments using prosodic features

Table 3 shows the effect of the respective features which are not informed about intonation boundaries (Table 3a) and those that are (Table 3b). Features that achieved a significant improvement over the baseline are marked in boldface.[10]

The best-performing feature in Table 3a is the presence of a pitch accent in short NPs. It can be seen that this feature has a negative effect when being applied on all NPs. Presumably, this is because the system is misled to classify a higher number of complex anaphoric expressions as non-anaphoric, due to the presence of pitch accents. This confirms our conjecture that long NPs will always contain *some* kind of accent and we cannot distinguish nuclear from prenuclear accents. Features based on GToBI(S) accent type did not result in any improvements.

(a) No boundary information

Baseline	47.93	
+ Feature applied to …	… short NPs only	… all NPs
PitchAccentType	45.31	46.23
PitchAccentPresence	**48.30**	46.57

(b) Including boundary information

Baseline	47.93	
+ Feature applied to …	… short NPs only	… all NPs
NuclearType (*n1* vs. *n2* vs. *pn* vs. *none*)	47.17	46.79
NuclearType (*n1*/*n2* vs. *pn* vs. *none*)	**48.55**	45.24
NuclearPresence (*n2*)	46.69	**48.88**
NuclearPresence (*n1*/*n2*)	**48.76**	47.47
NuclearBagOfAccents	46.09	**48.45**
NuclearFirst+Last	46.41	46.74

Table 3: CoNLL metric scores on DIRNDL for different prosodic features (no singletons, significant results in boldface)

Table 3b presents the performance of the features that are phonologically more informed. Distinguishing between prenuclear and nuclear accents *(NuclearType)* is a feature that works best for short NPs where there is only one accent, while having a negative effect on all NPs. Nuclear presence, however, works well for both versions (not distinguishing between *n1* or *n2* works for short NPs while *n2* accents only works best for all NPs). This feature achieves the overall best performance for both short NPs (48.76) and all NPs (48.88).

The *NuclearBagOfAccents* feature works quite well, too: this is a feature designed for NPs that have more than one accent and so it works best for complex NPs. Combining the features did not lead to any improvements.

Overall, it becomes clear that one has to be very careful in terms of how the prosodic information is used. In general, the presence of an accent works better than the distinction between certain accent types, and including intonation boundary information also contributes to the system's performance. When including this information, we can observe that when we look at the presence of a pitch accent (the best-performing feature), the distinction between prenuclear and nuclear is an important one: not distinguishing between prenuclear and nuclear deteriorates results. The results also seem to sug-

[6] `http://stel.ub.edu/semeval2010-coref/`

[7] Using the official CoNLL scorer v8.01, including singletons as they are part of TüBa 8

[8] Using the official CoNLL scorer v8.01, not including singletons as TüBa 9 does not contain them.

[9] CorZu performance: Don Tuggener, personal communication. We did not use CorZu for our experiments as the integration of prosodic information in a rule-based system is non-trivial.

[10] We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.01 level.

gest that simpler features (like the presence or absence of a certain type of pitch accent) work best for simple (i.e. short) phrases. For longer markables this effect turns into the negative. This probably means that simple features cannot do justice to the complex prosody of longer NPs, which gets blurred. The obvious solution is to define more complex features that approximate the rhythmic pattern (or even the prosodic contour) found on longer phrases, which however will require more data and, ideally, automatic prosodic annotation.

5 Conclusion

We have tested a set of features that include different levels of prosodic information and investigated which strategies can be successfully applied for coreference resolution. Our results on the basis of manual prosodic labelling show that including prosody improves performance. While information on pitch accent types does not seem beneficial, the presence of an accent is a helpful feature in a machine-learning setting. Including prosodic boundaries and determining whether the accent is the nuclear accent further increases results. We interpret this as a promising result, which motivates further research on the integration of coreference resolution and spoken language.

Acknowledgements

We would like to thank Anders Björkelund for his valuable comments on an earlier version of this paper, as well as Kerstin Eckart for her help with the preparation of DIRNDL data. This work was funded by the German Science Foundation (DFG), Sonderforschungsbereich 732 *Incremental Specification in Context*, Project A6, at the University of Stuttgart.

References

Marilisa Amoia, Kerstin Kunz, and Ekaterina Lapshinova-Koltunski. 2012. Coreference in spoken vs. written texts: a corpus-based analysis. In *Proceedings of LREC*, Istanbul.

Stefan Baumann and Arndt Riester. 2012. Referential and Lexical Givenness: semantic, prosodic and cognitive aspects. In Gorka Elordieta and Pilar Prieto, editors, *Prosody and Meaning*, pages 119–162. Mouton de Gruyter, Berlin.

Stefan Baumann and Arndt Riester. 2013. Coreference, Lexical Givenness and Prosody in German. *Lingua*, 136:16–37.

Stefan Baumann and Anna Roth. 2014. Prominence and coreference – On the perceptual relevance of F0 movement, duration and intensity. In *Proceedings of Speech Prosody*, pages 227–231, Dublin.

Stefan Baumann, Christine Röhr, and Martine Grice. 2015. Prosodische (De-)Kodierung des Informationsstatus im Deutschen. *Zeitschrift für Sprachwissenschaft*, 34(1):1–42.

Mary Beckman, Julia Hirschberg, and Stefanie Shattuck-Hufnagel. 2005. The original ToBI system and the evolution of the ToBI framework. In Sun-Ah Jun, editor, *Prosodic Typology – The Phonology of Intonation and Phrasing*, pages 9–54. Oxford University Press.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. In *Proceedings of LREC*, pages 3222–3228, Reykjavík.

Aoife Cahill and Arndt Riester. 2012. Automatically Acquiring Fine-Grained Information Status Distinctions. In *Proceedings of the 13th Annual SIGdial Meeting on Discourse and Dialog*, pages 232–236, Seoul.

Alan Cruttenden. 2006. The de-accenting of given information: a cognitive universal? In Giuliano Bernini and Marcia Schwartz, editors, *Pragmatic Organization of Discourse in the Languages of Europe*, pages 311–355. De Gruyter, Berlin.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A Discourse Information Radio News Database for Linguistic Analysis. In Sebastian Nordhoff Christian Chiarcos and Sebastian Hellmann, editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 65–76. Springer.

Caroline Féry. 1993. *German Intonational Patterns*. Niemeyer, Tübingen.

Carlos Gussenhoven. 1984. *On the Grammar and Semantics of Sentence Accents*. Foris, Dordrecht.

Manfred Klenner and Don Tuggener. 2011. An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *Proceedings of RANLP*, pages 178–185, Hissar, Bulgaria.

D. Robert Ladd. 2008. *Intonational Phonology (2^{nd} ed.)*. Cambridge University Press.

Jörg Mayer. 1995. Transcription of German Intonation. The Stuttgart System. University of Stuttgart.

Karin Naumann. 2006. Manual for the annotation of in-document referential relations. University of Tübingen.

Vincent Ng. 2010. 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.

Janet Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. 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.

Altaf Rahman and Vincent Ng. 2012. Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 798–807. Association for Computational Linguistics.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Stroudsburg, PA, USA.

Roger Schwarzschild. 1999. GIVENness, AvoidF, and Other Constraints on the Placement of Accent. *Natural Language Semantics*, 7(2):141–177.

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

Jacques Terken and Julia Hirschberg. 1994. Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech*, 37(2):125–145.

Joel Tetreault and James Allen. 2004. Dialogue structure and pronoun resolution. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium*, S. Miguel, Portugal.

Don Tuggener and Manfred Klenner. 2014. A hybrid entity-mention pronoun resolution model for german using markov logic networks. In *Proceedings of KONVENS 2014*, pages 21–29.

Carla Umbach. 2002. (De)accenting definite descriptions. *Theoretical Linguistics*, 2/3:251–280.