Annotating Discourse Anaphora

Stefanie Dipper
Institute of Linguistics
Bochum University
dipper@linguistics.rub.de

Heike Zinsmeister
Institute of Linguistics
Konstanz University
Heike.Zinsmeister@uni-konstanz.de

Abstract

In this paper, we present preliminary work on corpus-based anaphora resolution of discourse deixis in German. Our annotation guidelines provide linguistic tests for locating the antecedent, and for determining the semantic types of both the antecedent and the anaphor. The corpus consists of selected speaker turns from the Europarl corpus.

1 Introduction

An important component of text understanding is *anaphora resolution*, i.e. to determine the reference of constituents whose interpretation depends on (the reference of) other textual elements. The majority of anaphora are instances of noun phrase anaphora, which relate a noun phrase anaphor to a nominal antecedent. Grammatical restrictions (gender, number agreement) and saliency (grammatical function, recency) guide the resolution process in these cases. In addition to pronouns, definite noun phrases can be viewed as anaphoric in that they may corefer to some other NP in the given context. To solve the latter type of anaphora, lexical semantic knowledge is required, as provided by an ontology or a database like WordNet.

Another type of anaphora is discourse deixis (Webber 1988; 1991), which relates a noun phrase anaphor to a verbal or (multi-)clausal antecedent. The discourse entities that are introduced by antecedents of discourse deictic pronouns are called "abstract objects" since they refer to properties and propositional entities (Asher, 1993). Grammatical restrictions cannot apply since the antecedent is non-nominal and the anaphor—commonly in the form of a personal or demonstrative pronoun—is usually in neuter singular. We assume that in addition to saliency the resolution process needs to take semantic restrictions into account (cf. Hegarty et al. (2002)).

The automatic procedure of our research effort can be envisaged as follows: Given some text we first locate discourse anaphors. Next, the semantic (= abstract) type of each anaphor is determined, based on contextual features that are derived from annotated corpus data. The anaphor's semantic type restricts the semantic type of the antecedent, and thus narrows down the search space. Finally, the antecedent is located with the help of these semantic restrictions and, again, with contextual features derived from the corpus.

2 Related Work

Corpus-based studies have shown that abstract objects are less salient than other discourse referents, which has an effect on the choice of the anaphoric element (Hegarty et al., 2002). The abstract type of the antecedent and that of the anaphor do not necessarily coincide. The data suggests that reference to other types (referred to in the literature as coercion) is possible only in accordance to an abstractness hierarchy (Hegarty, 2003; Consten and Knees, 2005; Consten et al., 2007). The hierarchy starts with events as the most concrete type, which are anchored in spatial-temporal dimensions, and ends with propositions as the most abstract types. Anaphoric reference is possible to antecedents that are of the same type or less abstract than the anaphor (Consten and Knees, 2005).

Most works concerning the annotation of anaphora resolution do not make reference to abstract entities. OntoNotes, for example, only annotates reference to verbs (Pradhan et al., 2007). Annotation research efforts on discourse deixis include: Eckert and Strube (2000), Byron (2002), Poesio and Modjeska (2005), Poesio and Artstein (2008), and Müller (2007) for English; Navarretta (2000) for Danish; and Recasens (2008) for Spanish/Catalan. To our knowledge, there has been no attempt to systematically annotate such a corpus of German.

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 166–169,
Suntec, Singapore, 6-7 August 2009.

Test: Die Zusammenführung der nationalen und europäischen Ebene ist sehr wohl notwendig , obwohl natürlich die Haupttätigkeit in den Mitgliedstaaten stattfinden sollte und nur dann auf europäischer Ebene eingegriffen werden sollte , wenn **dies — *nämlich auf europäischer Ebene einzugreifen* —** unbedingt notwendig ist .

Anno: Die Zusammenführung der nationalen und europäischen Ebene ist sehr wohl notwendig , obwohl natürlich die Haupttätigkeit in den Mitgliedstaaten stattfinden sollte und nur dann [auf europäischer Ebene eingegriffen]$_{prop}$ werden sollte , wenn [**dies**]$_{prop}$ unbedingt notwendig ist .

Engl: 'It is indeed necessary to bring the national and European levels together, even though, of course, the main work should be done in the Member States, with the European level intervening only when **this** is absolutely necessary.'

Figure 1: Paraphrase test to determine the extension of the antecedent.

3 The Corpus

Our corpus consists of texts from the Europarl corpus (Koehn, 2005). As our basis, we selected all contributions whose original language is German (including Austrian German).

For the annotation task, we isolated medium-sized turns, consisting of 15–20 sentences. This was done to guarantee that the turns were not too lengthy but still provided enough information for the annotators to understand the broader context of discussion, so that they could resolve the anaphors without comprehension problems. From these turns, we selected those that contained the anaphor *dies* 'this'. This is the only anaphor in German which unambiguously refers to discourse units.

4 The Guidelines

Our guidelines are based on theoretical research on discourse semantics as well as work on annotating discourse phenomena.

Given some discourse anaphor (i.e., anaphoric *das, dies, was, es* 'that, this, which, it'), the guidelines define (i) how to locate the antecedent, (ii) how to determine the semantic type of the antecedent, and (iii) how to determine the semantic type of the anaphor. For each of these tasks, the guidelines provide linguistic tests (Dipper and Zinsmeister, 2009).

4.1 Locating the antecedent

To determine the antecedent of the anaphoric relation, a "paraphrase test" is applied: The annotator supplements the anaphor by a paraphrase in the form of *nämlich* … 'namely …'. The part that fills the … corresponds to the antecedent that we are looking for, cf. Fig. 1.[1] Antecedents can consist of VPs, (fragments of) main or subordinate clauses, or multiple sentences.[2]

4.2 The semantic type of the antecedent

We distinguish 10 types of propositional entities. Many verbs prototypically denote one type of propositional entity; *gewinnen* 'win', for instance, usually expresses an event. Often, however, the type of entity that is denoted depends on the context and usage of the verb; *Hans hat Äpfel gegessen* ('Hans ate apples') denotes a process, whereas *Hans hat zwei Äpfel gegessen* ('Hans ate two apples') denotes an event because the action has an end (when both apples are eaten)—i.e., the action is telic. The semantic types are defined in terms of the following features: world-dependent, time-dependent, dynamic, telic, and modal (with subtypes deontic and epistemic, generic, subjective) (see e.g., Vendler (1967), Asher (1993)). Table 1 displays the different types of propositional entities and their defining features. It also lists the labels used for annotating these entities. The entity types are ordered according to their degree of abstractness.

The entity type "deict" (deictic) does not fit in the abstractness hierarchy of the table. It refers to extra-linguistic entities, such as the external situation, or an issue that is currently the focus of attention in parliament, etc.

[1] The *Test* line displays the sentence with the anaphor (marked in bold-face) followed by the inserted paraphrase (in bold-face and italics). The *Anno* line shows the same example with the identified antecedent underlined. Both the antecedent and the anaphor are labeled with their semantic types (see below). The *Engl* line presents an English translation that is based on the original translations from Europarl. We used the tool OPUS (`http://urd.let.rug.nl/tiedeman/OPUS`) to retrieve the English translations.

[2] E.g., the anaphor *dies alles* 'all this' often refers to an antecedent consisting of multiple sentences. The actual antecedent can diverge from the one constructed by the paraphrase test in minor aspects, such as active-passive-alternations, or bare infinitive vs. *zu*-infinitive vs. participle. In some cases, the divergences are more important and could involve, for instance, the insertion or modification of the main verb. In such cases, annotators were asked to note and record the differences.

Prop. Entity		Label	Defining Features					Replacement Test
			W	T	Dyn	Tel	Mod	
1.	Event	ev	+	+	+	+	-	*Ereignis* ('event')
2.	Process	proc	+	+	+	-	-	*Vorgang* ('process')
3.	State	state	+	+	-	(-)	-	*Zustand* ('state')
4.	Circumstance	circ	+	+	-	-	-	*Umstand* ('circumstance')
5.	Modal (deontic + epistemic)	mod	+	+	-	-	mod	*Notwendigkeit, Möglichkeit, Chance, ...* ('necessity, possibility, opportunity, ...')
6.	Opinion, claim	op	+	+	-	-	subj	*Meinung, Ansicht, Behauptung, Einschätzung, Forderung, ...* ('opinion, view, claim, assessment, request, ...')
7.	Generic	gen	+	+/-	-	-	gen	*wohlbekannte, allgemeingültige Tatsache* ('the well-known, universal fact')
8.	Fact	fact	+	+/-	+/-	+/-	-	*Tatsache* ('fact')
9.	Proposition	prop	-	-	+/-	+/-	-	*(Art von) Aktivität, Aktion, Eigenschaft, ...* '(kind of) activity, action, property, ...')

Table 1: Semantic types and their defining features: W(orld), T(ime), Dyn(amic), (Tel)ic, Mod(al)

4.3 The semantic type of the anaphor

To determine the type of anaphors, we defined a "replacement test". With this test, the demonstrative anaphor *dies, das,* etc. is replaced by a suitable NP, such as *dieses Ereignis, dieser Vorgang*. The head noun indicates the type of the propositional entity (e.g., event, process).[3] Table 1 lists the different types of propositional entities and suitable replacement nouns. The annotators are asked to choose the *most concrete*, suitable noun.

5 Results

As a first pilot study on the reliability of our annotation guidelines, two student annotators annotated 32 texts that included 48 instances of the demonstrative pronoun *dies* 'this'. The pronouns were marked in bold face, and the annotation was performed on paper. After annotating 17 texts, the annotators discussed their intermediate results.

Locating the antecedent: In one case, one of the annotators decided on a deictic reading and did not mark an antecedent at all. 40 out of 47 antecedents (85%) were marked with identical spans. In four cases they chose differing but adjacent spans and in one case one of the annotators chose a longer string than the other.

The semantic type of the antecedent: The type of the antecedents coincided in 28 out of 47 cases (60%, α=0.52).[4] Agreement improved after the discussion period: 11/17 cases matched (α=0.60).

The semantic type of the anaphor: The results with respect to the semantic type of the anaphor seemed more disappointing: the annotators agreed in only 22 out of 48 instances (46%, α=0.37). However, after the discussion period, agreement leveled that of the type of the antecedent: 12 out of 17 cases coincided (α=0.66). In addition to the semantic type, we annotated the grammatical role of the anaphor, which occurred as the subject in 79% of cases and as objects elsewhere.

Annotators agreed most often on the four most concrete types ('ev, proc, state, circ') and least often on the three most abstract types ('gen, fact, prop'). This might be due to the fact that the most abstract types are applicable in many cases, but annotators are advised to choose the most concrete type that is available. In the majority of the cases (73%), the anaphor's type was identical with or more abstract than the antecedent's type.

6 Conclusion

In this paper, we presented a corpus-driven approach to discourse deictic anaphora in German. We introduced annotation guidelines that provide linguistic tests for locating the antecedent, and for determining the semantic types of both the antecedent and the anaphor. Further work will include exploitation of contextual information in combination with the semantic types to confine the set of potential antecedents.

Our corpus consists of selected speaker turns from the Europarl corpus. In this study, 32 texts (providing 48 instances of discourse deixis) were

[3] We use the term "semantic type of the anaphor" in a somewhat sloppy way. Put more precisely, the "semantic type of the anaphor" indicates the way that the anaphor refers to (parts of) the propositional discourse referent that is denoted by the antecedent.

[4] We computed α according to `www.asc.upenn.edu/usr/krippendorff/webreliability.doc`.

annotated according to these guidelines, and first results concerning inter-annotator agreement are promising (with an agreement of 85% on the extension of the antecedent, 60% on the antecedent type, and 46% on the type of the anaphor). The pilot study indicates that the paraphrase test helps the annotators in determining on the extension of the abstract antecedent.[5] It also shows that the linguistic tests for the semantic types have to be refined.

In the next steps, we will switch from paper-and-pencil annotation to annotation based on the tool MMAX2[6]. In addition to manually determining the semantic types of anaphors, we will investigate robust, fully-automatic approaches to the derivation of contextual features for anaphora resolution. For instance, we plan to take into account anaphors of the form *dieses Ereignis, dieser Umstand,* etc. ('this event, this circumstance'), which explicitly name the semantic type of the anaphor. In a later step other, more ambiguous, types of anaphors will be included in the investigation.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments, and our student annotators: Doris Aimer, Iris Bräuning, Christine Enzinger, Stefanie Konetzka, Barbara Mrzyglod. The work was in part supported by Europäischer Sozialfonds in Baden-Württemberg.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Boston MA.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the ACL-02 conference*, pages 80–87.

Manfred Consten and Mareile Knees. 2005. Complex anaphors — ontology and resolution. In P. Dekker, editor, *Proceedings of the 15th Amsterdam Colloquium*, Amsterdam: University.

Manfred Consten, Mareile Knees, and Monika Schwarz-Friesel. 2007. The function of complex anaphors in texts: Evidence from corpus studies and ontological considerations. In *Anaphors in Text*, pages 81–102. John Benjamins, Amsterdam/Philadephia.

Stefanie Dipper and Heike Zinsmeister. 2009. Annotation guidelines "Discourse-Deictic Anaphora". Draft. Universities of Bochum and Konstanz.

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

Michael Hegarty, Jeanette K. Gundel, and Kaja Borthen. 2002. Information structure and the accessibility of clausally introduced referents. *Theoretical Linguistics*, 27(2-3):163–186.

Michael Hegarty. 2003. Type shifting of Entities in Discourse. Presentation at the First International Workshop on Current Research in the Semantics-Pragmatics Interface, Michigan State University.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Christoph Müller. 2007. Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 816–823, Prague, Czech Republic, June.

Costanza Navarretta. 2000. Abstract Anaphora Resolution in Danish. In *1st SIGdial Workshop on Discourse and Dialogue*, pages 56–65, Hong Kong, China, October.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the LREC 2008*, Marrakech, Morocco.

Massimo Poesio and Natalia N. Modjeska. 2005. Focus, activation, and *this*-noun phrases: An empirical study. In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing*, volume 263 of *Current Issues in Linguistic Theory*, pages 429–442. John Benjamins.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*, Irvine, CA.

Marta Recasens. 2008. Discourse deixis and coreference: Evidence from AnCora. In *Proceedings of the Second Workshop on Anaphora Resolution (WAR II)*, pages 73–82.

Zeno Vendler, 1967. *Linguistics in Philosophy*, chapter Verbs and Time, pages 97–121. Cornell University Press, Ithaca.

Bonnie L. Webber. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of the ACL-88 conference*, pages 113–122.

Bonnie L. Webber. 1991. Structure and ostention in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6:107–135.

[5] The study was restricted to instances of the unambiguous anaphor *dies* 'this', which might have simplified the task of selecting an antecedent.

[6] MMAX2: `http://mmax2.sourceforge.net/`.