Towards the Automatic Resolution of Anaphora with Non-nominal Antecedents: Insights from Annotation

Adam Roussel, Stefanie Dipper
Ruhr-Universität Bochum
Fakultät für Philologie
Universitätsstraße 150
44780 Bochum
{roussel, dipper}
@linguistics.rub.de

Sarah Jablotschkin, Heike Zinsmeister
Universität Hamburg
Fakultät für Geisteswissenschaften
Überseering 35, Postfach #15
22297 Hamburg
{sarah.jablotschkin,
heike.zinsmeister}@uni-hamburg.de

Abstract

This paper deals with a particular form of anaphora in which the anaphors refer to non-nominal antecedents. We investigate two existing datasets, annotated with pronominal and nominal anaphors (shell nouns) respectively, and attempt to determine to what degree the different types of anaphors provide useful hints as to the form and location of their antecedents. To this end, we look at the distribution of the antecedents, their syntactic form, and their semantic content. In particular, as the difficulty of annotating the phenomenon constitutes a major hurdle to the development of larger datasets, we take a close look at the agreement between annotators and relate this to the different types of anaphors.

1 Introduction

Coreference and anaphora resolution constitute a fundamental part of natural language analysis and understanding and provide valuable information for high-level tasks, such as automatic text summarization, machine translation, or question answering.

A particularly challenging form of anaphora are those cases in which the anaphor refers to an abstract entity, such as an event or a fact, and the antecedent is typically expressed by a complex constituent, i.e., a clause or sentence. In (1a) from Eckert and Strube (2000), the anaphor is *this* and refers to the event of John crashing the car, which is expressed by the preceding clause.[1] Such anaphors can be either neuter pronouns (*it, this, that*) or noun phrases headed by certain abstract nouns, e.g., *problem* or *fact*, called shell nouns (Schmid, 2000), as in (1b).

(1) <u>John crashed the car.</u>
  a. **This** shows how careless he is.
  b. **This fact** shows how careless he is.

Anaphora resolution for anaphors with nominal antecedents usually makes use of morpho-syntactic properties such as gender and number, and the anaphor and its antecedent are assumed to agree with regard to these features. Anaphora with non-nominal antecedents, however, cannot refer to such properties, since clauses do not have gender or number features. Moreover, locating the antecedent is considerably more difficult because non-nominal antecedents can be of various syntactic types, e.g., simple verb phrases, complex sentences, or discontinuous constituents. It can even be the case that the anaphor's referent is not overtly expressed at all. Accordingly, current resolution systems only handle a subset of such anaphora or do not attempt to process them at all.

OntoNotes (Hovy et al., 2006) addresses the problem by marking verbal heads only, leaving the exact boundaries of the antecedent unspecified. In general, it is unclear what information about the anaphor's referent needs to be retrieved from the antecedent string. The answer, of course, largely depends on the task at hand, and for some applications the OntoNotes solution might suffice. Full natural language understanding, however, would need more detailed information about the referent.

In this paper, we investigate to what degree antecedents exhibit anaphor-specific properties, or, in other words, whether the different types of anaphors (pronouns vs. shell nouns) provide useful hints as to what their antecedents may look like. To this end, we look at the distribution of the antecedents (in terms of direction and distance), their syntactic form (e.g., infinitive vs. finite clause), and their semantic content.

In particular, we consider and examine annotation difficulty (how and to what extent annotators

[1] In the examples, anaphors will be set in boldface and antecedents underlined.

disagree) and relate the results to the different types of anaphors we observe. That is, this paper provides an in-depth anaphor-specific evaluation of our previous annotation efforts.

The paper is structured as follows. Section 2 addresses the different terms used for the phenomenon of anaphora with non-nominal antecedents and presents related work. Section 3 and Section 4 introduce the datasets with pronominal anaphors and shell nouns, respectively, and summarize previous results. We extend these previous analyses by investigating differences between different types of anaphors (Section 5) and between individual annotators (Section 6), and Section 7 concludes the paper.

2 Terminology and Related Work

The phenomenon that we address in this paper has been discussed in the literature under a variety of different terms. For instance, Webber (1988), Eckert and Strube (2000), Byron (2004), and Recasens (2008) use the term *discourse deixis*. This term emphasizes the fact that the anaphor points ("deictically") to some part of the discourse model, which serves as the referent of the anaphor.[2]

Another term is *abstract anaphora*, as used, e.g., by Asher (1993), Navarretta (2007), and Dipper et al. (2011). The focus here is on the fact that such anaphora tend to refer to some abstract entity, such as an event or fact. The term *complex anaphora* (Consten et al., 2007) hints at the complex nature of such entities, which can involve multiple participants and the relations between them.

In this paper, we adopt the term *non-Nominal-Antecedent (non-NA) anaphora* (Kolhatkar et al., 2018), emphasizing the syntactic aspect of the phenomenon, since we consider the antecedent's non-nominal syntactic type both a characteristic property of this type of anaphora and a special challenge for annotation and automatic resolution. We define the *antecedent* as the string in the anaphor's context which most closely represents the referent of the anaphor. The *semantic type* of the referent classifies it, e.g., as an EVENT, STATE, FACT, or PROPOSITION (see Asher (1993)).

Some authors argue that non-NA anaphora referring to a certain abstract referent may require *referent coercion* in which an abstract referent is converted into the semantic type the anaphor requires (e.g., Byron (2004)). For instance, in (2) from Eckert and Strube (2000), whereas the antecedent *John crashed the car* might ordinarily be construed as an EVENT, the anaphor *this* refers to it as a FACT, which causes it to be interpreted as such. Consten et al. (2007) provide similar examples for referent coercion with shell noun anaphora, such as (3), in which the anaphor *this fact* refers to an EVENT antecedent.[3]

(2) John crashed the car. **This** shows how careless he is.

(3) The Americans tried to invade the building but were forced back by shots from the top floor. Rumsfeld had to explain the consequences resulting from **this fact** during a press conference in the afternoon.

Most of the work on non-NA anaphora has been done for English. Early analyses and annotation efforts include Schiffman (1985), Passonneau (1989) and Webber (1988), who focus on the pronominal anaphors *it*, *this*, and *that*. Francis (1986) and Schmid (2000) were among the first to examine shell nouns in English in more detail. A comprehensive survey for English is provided by Kolhatkar et al. (2018). Prominent work for languages other than English include Recasens (2008) on Spanish and Catalan and Böhmová et al. (2003) on Czech, among others.

Most of these works focus on either pronominal anaphora or shell nouns, and only a few consider both, e.g., Recasens (2008) or Uryupina et al. (2018). To our knowledge, none of them investigates and compares annotation results for both.

3 Pronominal Anaphors

The data used in both settings (pronouns and shell nouns) come from the Europarl Corpus (Koehn, 2005). In a first project, extensive annotation guidelines were created and applied to pronominal anaphors in roughly 100 German and English turns (i.e., contributions of German and English speakers in the European Parliament). In a second project (described in Section 4), the guidelines were slightly adapted and applied to shell nouns instead. Both projects used the annotation software MMAX2 (Müller and Strube, 2006). In the current paper, only the German subcorpora are considered.

[2] *Discourse deixis* is also sometimes used to refer to the linguistic form of an utterance, e.g., with expressions like *this sentence* , which we classify as ordinary deixis.

[3] Consten et al. (2007) also discuss German examples from the Tiger corpus (Brants et al., 2004).

This section briefly describes the first project, pertaining to pronominal anaphors, and summarizes results from previous studies, reported in detail in Dipper et al. (2011), Dipper and Zinsmeister (2012), and Zinsmeister et al. (2012).[4]

3.1 Annotation project

The dataset *non-NA-pro* contains 100 turns with 643 anaphor candidates in the form of pronominal *dies* 'this', *das* 'that', and *es* 'it', which were automatically highlighted for the annotators.[5] The annotations were created by four different annotators in several consecutive annotation steps:[6]

(i) Identify non-NA anaphors: Exclude anaphors with nominal antecedents and expletive *es* 'it'.

(ii) Identify the non-nominal antecedent: Find a string in the context that best represents the content of the anaphor. To this end, a linguistic test is applied in which an appositive *namely* phrase is appended to the anaphor. (4a) illustrates this: The anaphor is *this* (in boldface), and the phrase between the dashes (in italics) has been added. It picks up a string from the prior context (underlined in the example), which is to be marked as the antecedent.

Often, a string cannot be used literally but only in modified form. If such modifications concern content words rather than just functional words or inflectional markings, as in (4a), the string is marked as "divergent".

(iii) Determine the semantic type of the anaphor: Select a noun from a predefined list (containing nouns like *Umstand* 'circumstance' or *Entwicklung* 'development') that could replace the anaphor and fits semantically. The choice in (4b) shows that the anaphor refers to an abstract entity of the type MEASURE.

(4) a. . . . with the European level intervening only when **this**— *namely that the European level intervenes*—is absolutely necessary.

b. . . . with the European level intervening only when **this measure** is absolutely necessary.

3.2 Previous results

We start by summarizing inter-annotator agreement of each step, as described in Dipper and Zinsmeister (2012): In step (i), 225 out of 643 candidates were identified as non-NA anaphors by the annotators, with an inter-annotator agreement of Cohen's $\kappa = .79$. In step (ii), strings with an average length of 13.9 tokens ($\sigma = 10.6$) were marked as antecedents. These strings could be freely marked and the authors did not compute expected agreement.[7] They calculated observed agreement (A_o) instead: If only exact matches are considered, $A_o = 0.40$, if matching head verbs are considered, $A_o = 0.55$, and if overlapping antecedents are considered, $A_o = 0.84$. In step (iii), annotators could specify up to two nouns. If they agreed on any of them, it was considered a match. Here, $A_o = 0.75$.

Apart from calculating the inter-annotator agreement for each of the different annotation steps, Dipper and Zinsmeister (2012) found that *dies* 'this' is rare in general, and the most frequent pronominal non-NA anaphor is *das* 'that'. The majority of *das* and *dies* usages are instances of non-NA anaphors, in contrast to *es* 'it', which is only rarely used as a non-NA anaphor. Antecedents tend to be rather short when *es* is the non-NA anaphor, and somewhat longer when it is either *das* or *dies*. The study also investigated distance (measured in number of tokens) between the non-NA anaphor and its antecedent. It turns out that *das* most often occurs in close proximity to its antecedent while *dies* tends to occur at a greater distance from its antecedent. In contrast to *das* and *dies*, *es* is also used as a cataphor, i.e., preceding its 'antecedent'.

Dipper et al. (2011) and Zinsmeister et al. (2012) investigated further syntactic properties of non-NA anaphors. They found that non-NA anaphors predominantly function as the subject of a sentence and less often as the object.[8] Grammatical function

[4] Note that the studies differ slightly with regard to their datasets: Dipper et al. (2011) and Zinsmeister et al. (2012) report on a dataset of 94 turns with 225 (Dipper et al., 2011) and 203 (Zinsmeister et al., 2012) pronominal non-NA anaphors, respectively; Dipper and Zinsmeister (2012) report on a superset of 100 turns with 223 pronominal non-NA anaphors. This is due to the fact that the annotations were produced over several years, and were filtered (anaphors) and enlarged (turns) during that process.

[5] Instances of *was* 'which' were also automatically highlighted. However, in the present corpus, none of them were used as a non-NA anaphor, and the annotators excluded them from further annotation.

[6] An additional step (determining the semantic type of the antecedent) is omitted here since we do not use these annotations in the current study.

[7] Note that Krippendorff's unitized α_u (Krippendorff, 1995; Krippendorff, 2013) would allow doing that.

[8] The annotations only cover the nominative/accusative forms *das*, *dies*, and *es*. Non-NA anaphors functioning as prepositional objects or adjuncts would not be realized, e.g.,

correlates with syntactic position. Accordingly, non-NA anaphors are located most often in the sentence-initial *prefield* position—a position that is very often (but not always) filled by the subject in German. Some non-NA anaphors occur embedded in a subordinate clause, but relatively few occur in the matrix clause in a position other than in the prefield.

4 Shell Nouns

This section describes the shell noun annotation project and summarizes the results from Simonjetz and Roussel (2016).

4.1 Annotation project

The dataset *non-NA-sn* contains annotations obtained in the course of the study described in Simonjetz and Roussel (2016), which aimed to examine and compare the realization of shell nouns in German and English.

In all, 371 turns were selected for annotation by two expert annotators (the study's authors), who annotated the English and German translations of these turns in parallel. The authors aimed to cover a greater variety of shell noun related phenomena than had been discussed previously. Whereas Schmid (2000) focused on shell noun instances which were discoverable by means of lexico-syntactic patterns, this study was intended to cover anaphoric and cataphoric shell nouns, singular and plural shell nouns, and coordinated shell nouns, as well as shell nouns with nominalized antecedents or multiple coordinated antecedent phrases. For this reason, the study required annotators familiar with the theoretical background of the phenomenon.

A list of 50 nouns in each language, which were determined to be generally frequent and to occur frequently in typical shell noun patterns (using the statistics from Schmid (2000) for English and Simonjetz (2015) for German), were selected as the set of nouns to be annotated. This was intended both to help the annotators annotate as many instances as possible and to keep the annotations easily comparable between annotators.

The annotators were presented with a series of Europarl turns as English–German pairs in parallel. For each pair of turns, the annotators examined each such pre-selected noun, and carried out the following annotation instructions:

(i) Decide whether or not this instance is functioning as a shell noun.

(ii) If yes, then select the span of tokens that best represents the content of this shell noun. Create a pointer from the shell noun to its antecedent.

(iii) Align every candidate shell noun instance with the best matching counterpart in the parallel turn in the other language.

(iv) If some antecedent was assigned, align this with the best matching span in the other language.

4.2 Previous results

For the annotation of the shell noun instances themselves, i.e., whether an instance is or is not a shell noun instance, the authors calculated an agreement value of Cohen's $\kappa = 0.73$ (Artstein and Poesio, 2008); $A_o = 0.86$.

Comparing the spans identified as antecedents is more complicated, since it is not clear how to weight disparities in which tokens are marked and indeed which spans should be considered comparable at all. The authors calculate Krippendorff's unitizing α_u, which had been used in similar studies for this purpose (Kolhatkar and Hirst, 2012), and arrive at a value of $\alpha_u = 0.84$ for the agreement between the annotators for all antecedent spans, indicating relatively good agreement overall.

The analysis in Simonjetz and Roussel (2016) showed that German shell nouns tend to have greater variation in the locations of their antecedents than in English, in which case the antecedent usually follows directly after the shell noun, most likely as a subordinate clause. This tendency suggests that string-based search methods, which are useful for discovering English shell nouns, are unlikely to be as useful for other languages. The authors also noted that German shell nouns seem to refer to nominalized antecedents (i.e., NPs headed by a deverbal noun, as in *Aktualisierung der Software* 'updating the software') more often.

5 Further Analysis

In this section, we present a more detailed analysis of the gold-standard subsets of the non-NA-pro

as *auf das/es* 'on that/it', but rather with a pronominal adverb, such as *darauf* 'thereon'.

and non-NA-sn datasets described above. We attempt to determine whether the use of a particular pronominal non-NA anaphor, e.g., the pronoun *this*, or a particular shell noun, such as *Umstand* 'circumstance', could tell a resolution system where to seek resolution candidates and what surface form the best candidates are likely to have.

5.1 Conceptualizing pronominal anaphors

Whereas pronouns like *dies*, *das* and *es* are inherently vague and ambiguous, the semantic type of a shell noun anaphor is made explicit (or at least constrained) by the shell noun itself. By evaluating which nouns are chosen to replace the non-NA anaphoric pronouns in the annotation task (cf. Section 3.1), it is possible to gain insights into how the annotators conceptualize non-NA anaphoric pronouns in context and to find out whether the pronouns are completely interchangeable or whether they differ with respect to their semantics.

Figure 1 is based on a gold version of the non-NA-pro annotation task and illustrates how the different pronouns are conceptualized in terms of replacement nouns (cf. Section 3.1).

Clearly, there are differences between the semantic contexts that relate to the respective pronouns. There are a number of nouns that seem to go with all three different pronouns, such as *Maßnahme* 'measure' and *Sachverhalt* 'circumstance', and others that typically go with just one or two of the pronouns. For example, *Umstand* 'circumstance' has a strong tendency to be picked to replace *das*. Some replacement nouns are even exclusively used for *das* 'that'. These are *Ansicht* 'view', *Meinung*'opinion', *Frage* 'question', *Notwendigkeit* 'necessity' and *Ereignis* 'event'. *Aktivität* 'activity', on the other hand, is a noun that typically replaces *es* 'it'. There are no nouns that only go with *dies* 'this', which may indicate that *dies* has even less specific semantics than *das* and *es*.

5.2 Antecedent syntactic types

Table 2 in the appendix shows the relative frequency with which the various pronouns and shell noun lemmas were annotated together with a particular antecedent type. The antecedent types were determined automatically from the tags and dependency parses produced with the `spaCy` toolkit (Honnibal and Montani, 2017). For instance, if the antecedent's string ends with a question mark, the type is determined as QUEST; see Table 3 in the appendix for a list of all types.

As in Asher (1993), we anticipate that the surface form of the antecedent and a lemma's preference for antecedents of a particular syntactic type can be clues to the semantic type of those non-NA anaphors. Such types could be a valuable source of information for systems attempting to resolve non-NA anaphora. Given a particular shell noun, such as *Absicht*, we would expect that it strongly prefers a ZU-type antecedent and could constrain the search for antecedents accordingly.

Indeed, our annotation results (Table 2) show that the shell nouns *Absicht* 'intent' and *Plan* 'plan' seem to prefer ZU-type antecedents, which coincides with our intuition that these nouns describe actions to be taken. Likewise the strong preferences of *Tatsache* 'fact' and *Meinung* 'opinion' for DASS-type antecedents can be seen as reflecting the propositional nature of their referents.

Similarly, *es* 'it' is often associated with events and actions,[9] and yet here (Figure 2), *es* does not seem to have any particular preference towards V-FIN or ZU-type antecedents, which are usually associated with these semantic types.

5.3 Antecedent distribution

Figure 3 in the appendix gives an overview of the distances between the annotated shell nouns and their antecedents, which are grouped according to the shell noun lemmas involved. Some of the lemmas seem to be used much more frequently with content which was mentioned previously (anaphora), whereas others seem to be preferentially used to introduce content which comes later in the text or in a subordinate clause.

We also see that certain shell nouns show much more variation in the locations of their antecedents than others. Shell nouns with especially distant antecedents appear to be ones that could equally refer to the text itself (ordinary deixis) as well as to its content (discourse deixis). This would include such nouns as *Frage* 'question/matter' and *Argument* 'argument'.

Note that this diagram does not contain frequency information: Most of the individual shell noun lemmas are relatively infrequent in the data, such that any generalizations we make here must remain tentative until more data have been collected.

[9] As in this example from Byron (2002):

(i) Each fall, penguins migrate to Fiji.

a. *That*'s why I'm going there next month. (FACT)

b. *It* happens just before the eggs hatch. (EVENT)

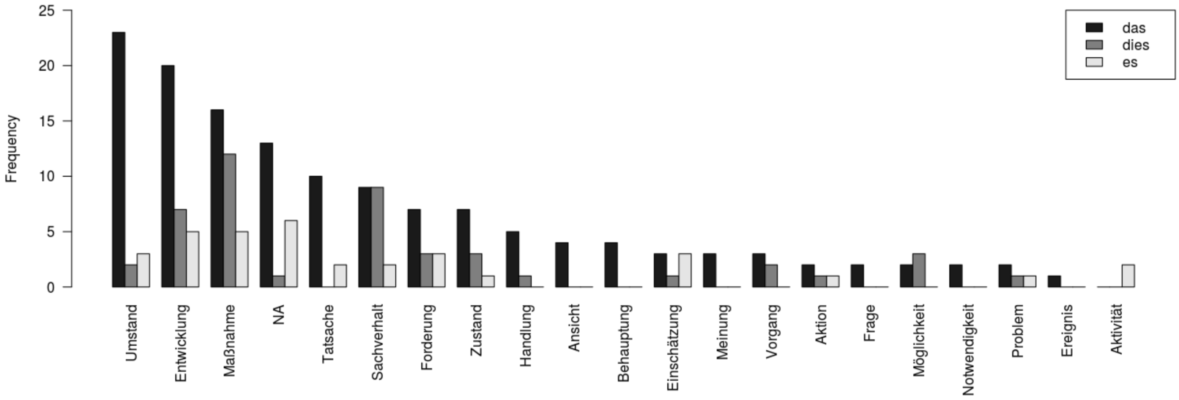

Figure 1: Replacement nouns by pronoun (n = 183)

6 Comparing Annotators

It is important to not just resolve or otherwise discard instances on which annotators disagree, since these mismatches can point to interesting properties of the annotation instances. In this section, we examine more closely the data of the annotation projects under consideration by comparing the results obtained by two individual annotators on the same data. We will analyze the disagreements in the annotations, which are provided in the publicly available annotated data[10] and which have not—to our knowledge—been analyzed in detail before. We restrict this to instances that have been identified as instances of non-NA anaphora by both annotators before consolidation ($n_{\text{non-NA-pro}} = 183$, $n_{\text{non-NA-sn}} = 334$).

Our analyses are guided by such questions as: Did the annotators agree more often on, for instance, *Tatsache* 'fact' than on *Frage* 'question'? If so, can we identify a property of *Frage* that explains why it is more challenging to decide on a common antecedent for *Frage* than for *Tatsache*?

6.1 Non-NA anaphoric status

The first step in both of the annotation tasks was to decide whether the preselected instances of pronouns and shell nouns were non-NA anaphors or not. Overall agreement results have been presented in Section 3.2 and Section 4.2 above.

Lemma-specific evaluations show that non-NA anaphoric status was most reliably identified with *dies* 'this' ($A_o = 0.81$), while *es* 'it' ($A_o = 0.76$) and *das* 'that' ($A_o = 0.73$) were harder to classify. With regard to shell nouns, Table 4 in the appendix shows that for lemmas with frequencies > 5 observed agreement in general is rather high. Agreement was perfect with nine lemmas, and above 0.8 with 21 lemmas. It is not clear what makes these nine lemmas particularly easy to identify. In some cases, there might be a relation to the syntactic form of the antecedent (see Section 5.2). For instance, *Ansicht* 'opinion', *Hinweis* 'hint', *Tatsache* 'fact', and *Überzeugung* 'belief' occur predominantly with DASS-type antecedents. Other lemmas like *Annahme* 'assumption' and *Wille* 'will' were not used at all as anaphoric shell nouns in our data.

6.2 Replacement nouns

Figure 2 illustrates how the annotators agreed or disagreed in their choice of replacement nouns for particular instances of non-NA anaphoric pronouns (cf. step (iii), Section 3.1).

Relatively few nouns were points of agreement for the two annotators: *Entwicklung* 'development', *Maßnahme* 'measure', and *Umstand* 'circumstance' were among the nouns for which there was the most consensus.

However, the disagreement on the replacement nouns is not arbitrary, as the mismatches often concern semantically similar replacement nouns. For example, when annotator 2 chose *Zustand*, annotator 1 might choose a semantically closely-related noun, such as *Umstand* or *Sachverhalt*—all of these nouns refer to some kind of state or situation. When annotator 1 chose *Einschätzung*, annotator 2 sometimes chose *Forderung* or *Meinung*, all of which refer to the speaker's opinion or attitude. Higher agreement scores might have been achieved

[10] https://github.com/ajroussel/nna-datasets

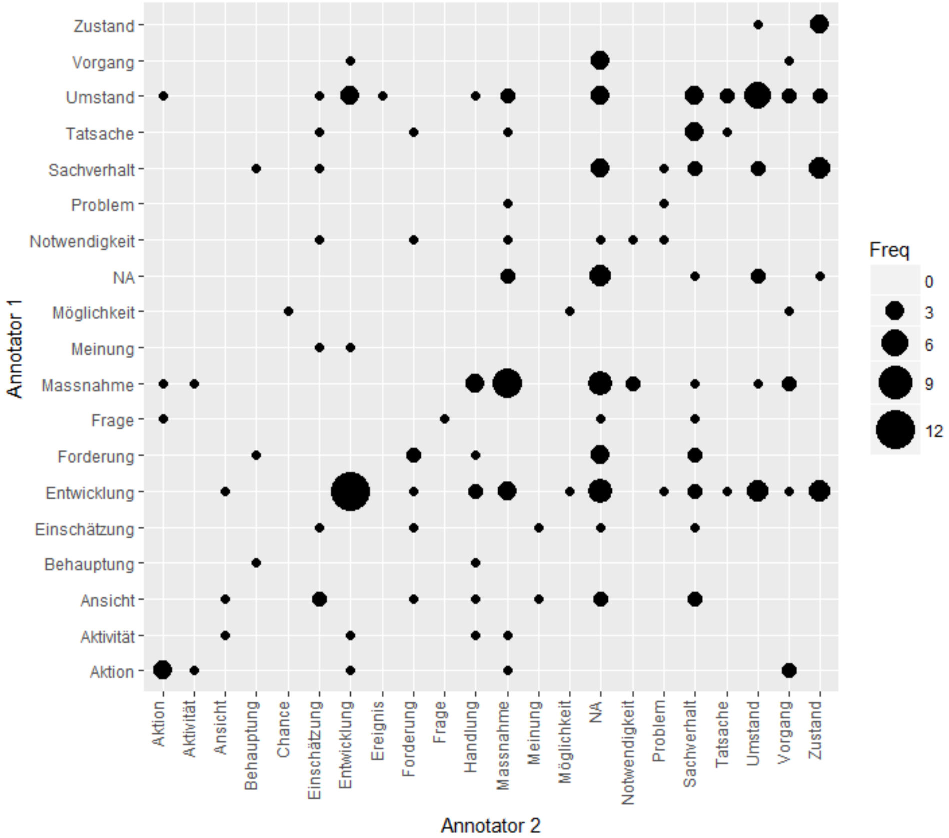

Figure 2: Comparison of replacement nouns selected by annotators

if the annotators had been given fewer but semantically more distinctive replacement nouns to choose from.

6.3 Antecedent strings

Clearly delimiting the antecedents of non-NA anaphora is a non-trivial task. This can be illustrated by comparing the antecedents of pronominal non-NA anaphors as annotated by two annotators.

Our first finding on the pronominal set is that in the majority of cases the annotators disagree by at least one token on the boundaries of the antecedent token sequence. Only 26% of cases were marked identically by both of the annotators. However, of the remaining 74%, the vast majority do involve some degree of overlap. Just 11% of the instances involve no overlap at all.

Table 1 shows the degree to which agreement differs between the pronouns: Agreement is considerably higher for *es* ‘it’ than for *das* ‘that’ or *dies* ‘this’. Not only does *es* generally have shorter antecedents than the other pronouns (cf. Section 3.2), its antecedents also seem to be more easily identifiable. These distinctive characteristics of *es* in anaphoric constructions in comparison to the other two pronouns might have something to do with the pronoun’s lexical features: In contrast to *das* and *dies*, *es* cannot be used as a demonstrative pronoun.

Pronoun	Identical
das	0.21
dies	0.30
es	0.48

Table 1: Proportion of identically annotated antecedents per pronoun (n = 183).

For the shell noun antecedents, the percentage of overall agreement was considerably higher: In 87% of the shell noun instances identified by both annotators (n = 334), the antecedents were identical. This accords with our intuition that shell noun content tends to be located in more predictable environments.

6.4 Antecedent mismatch types

In order to examine the nature of the divergences more closely, a meta-annotation was carried out by two of the authors: One or more of several syntactic categories were assigned to spans that differ between annotators. For instance, if one of the annotators included the subject in the antecedent and the other did not, this is assigned the category SUBJ.[11] The results of the meta-annotation are displayed in Table 5 in the appendix. The numbers are the relative frequencies of divergent antecedents in the respective category in relation to the total numbers of antecedents marked by both annotators (n = 183 for pronouns, n = 334 for shell nouns). Note that the columns do not add up to one because an antecedent pair can have multiple labels, e.g., the antecedent of one annotator could be missing the finite auxiliary verb (A-FIN) and the subject (SUBJ) with respect to the antecedent marked by the other annotator.

The distributions of disagreement in marking the antecedents of pronouns and of shell nouns show a moderate, but significant, correlation (Spearman's $\rho_S = 0.5408845$, $p = 0.0458$, calculated without PUNCT), which suggests that regardless of whether pronouns or shell nouns are being annotated, the divergent spans between annotators can be described in terms of the same syntactic categories.

6.4.1 Pronoun antecedents

For pronoun antecedents, the largest category of divergences (29%) relates to whether a punctuation mark was included in the antecedent string or not (category PUNCT). These cases obviously do not represent real divergences and could be easily avoided with more explicit annotation instructions regarding punctuation marks.

The second-largest group of divergences (14%) pertains to the marking of subjects (category SUBJ). This category not only includes cases in which one annotator marked a subject and the other did not, but also a number of cases in which the annotators marked different subjects: Where one annotator marked a pronoun, the other might resolve this anaphor and mark a coreferent NP instead. For example, in (5), annotator 1 marked *das Programm* 'the program' as part of the antecedent while annotator 2 marked the coreferent *es* 'it' instead.[12]

(5) Wir sind sehr glücklich über das Programm "Jugend in Aktion", weil es eine Vielzahl von Anregungen des Rechnungshofes – auch aus den vergangenen Jahren – berücksichtigt und ganz entschieden auch Programmvereinfachung auf seine Fahnen geschrieben hat. **Dies** ist ein höchst interessanter Ansatzpunkt. . . .

Proportionally, there were more shell noun antecedents with no overlap at all (26%) than there were pronoun antecedents in this category (11%) (category ALL). However, as pointed out above, agreement on antecedents between annotators was much higher overall with shell nouns than with pronouns. It seems that, most of the time, shell noun antecedents are easier to identify, probably because of the syntactic environments that are easier to predict than those of antecedents of pronouns. But when they don't occur in these contexts, they are all the more ambiguous, and it is more likely that antecedent annotations will diverge completely.

In the case of non-overlapping antecedents, there were several instances in which the annotators marked complementary clauses of one sentence. In (6), one of the annotators marked the main clause (*Ich meine* 'I think') as the antecedent, the other the subordinate clause.

(6) Ich meine, [. . .], dass ein Kompromiss gefunden werden musste und dass dies unter den obwaltenden Umständen ein fairer Kompromiss ist, dass wir als Europäische Gemeinschaft aber auch dafür Sorge tragen müssen, dass die Menschen Vertrauen zu diesem Plan fassen, und dass wir dafür werben sollten, dass dieses Vertrauen auch von unserer Seite untermauert sein muss.

6.4.2 Shell noun antecedents

With shell noun antecedents, on the other hand, there were a high percentage of divergences in the category SENT (31%). That is, in a coordination of sentences, such as in (7), one annotator marked only one conjunct as the antecedent whereas the other one marked both.

(7) Vor noch gar nicht allzu langer Zeit wurde hierzu eine Einigung erzielt, aber das war schon vor sieben oder acht Jahren, und damit hatten die Duty-Free-Läden die **Möglichkeit**, die Auswirkungen auf ihre Geschäftstätigkeit zu prüfen und nach Alternativen zu suchen.

[11] We also annotated whether one of the annotators marked a nominal antecedent (category NP in Table 5), which was part of the shell noun annotation task but not the pronoun task. Despite the instructions, this was the case for 3% of the non-NA anaphoric pronouns identified by both annotators.

[12] In the examples, spans marked by one annotator only are underlined with wavy lines or dots, and spans marked by both annotators are double-underlined.

Also, 12% of the shell noun antecedents differed because one of the annotators marked a complex sentence whereas the other only marked the matrix clause (category C-SUB). These findings indicate that often both annotators identify some sentence-like structure as the antecedent but disagree on its exact delimitation.

We close with some observations that do not correspond to distinct annotation categories but were striking nonetheless.

Punctuation marks (e.g., colons, dashes) followed by an extraposed part of the sentence seem to be likely to trigger an antecedent mismatch. For example in (8), annotator 2 included *: die Sozialpolitik.* 'the social policy' in their antecedent string, but annotator 1 did not.

(8) Der jetzt zwischen den großen Fraktionen kursierende Kompromiss hat einen äußerst verschwommenen Begriff eingeführt : die Sozialpolitik. **Das** ist höchst bedauerlich ...

The individual finite verb categories (V-FIN, M-FIN, etc.) did not yield very high proportions, but taken together, 12% of the pronoun antecedents differed because one annotator marked a finite verb and the other did not, as in (9). It might be the case that annotators tend to include the finite verb because it is the head of the sentence and they think the antecedent should be as complete a syntactic structure as possible—and not because they think that the antecedent with the verb better represents the content of the anaphor.

(9) Aber das Projekt muss entschlossen, gemeinschaftsorientiert und visionär zur politischen Union weiterentwickelt werden. Wenn **dies** nicht geschieht, verlieren wir das Vertrauen der Bürger.

This impression is reinforced by the fact that in the meta-annotation the finite verb categories often occur together with elements in the prefield, like adverbs or prepositional phrases (category ADV), or subjects (category SUBJ).

7 Improving Non-NA Anaphora Resolution: Concluding Remarks

The resolution of non-NA anaphora remains a challenging task for computational systems. In this paper, we re-examined the process and the results of previous annotation studies addressing this topic hoping to uncover useful information that can inform the design of systems in the future.

One way we might use this information would be to address problems relating to data and annotation. With a better understanding of the ways in which annotators disagree, we can leverage this information to learn more about non-NA anaphora in general. That inter-annotator agreement was worse for *Anstrengung* 'exertion' than for *Bemühung* 'effort' indicates that even shell nouns with similar meanings can vary with respect to the ease with which they are resolved. Future research could look into the (possibly semantic) factors behind this tendency. The disagreements also say something about how well the annotation scheme represents the phenomenon being annotated, i.e., whether or not the labels provided are adequate. Other disagreements (such as whether or not punctuation was included) may be less informative but can point us to better approaches to annotation that can minimize such errors.

Second, our analysis has hopefully uncovered tendencies in the data that can assist in the design of machine learning features for recognizing or resolving non-NA anaphora. Looking closer we might find classes of shell nouns that behave similarly, e.g., are usually used anaphorically or whose antecedents occur roughly 10 tokens away. Then by comparing novel shell noun candidates semantically to these known instances we might also have some idea where we are most likely to find their antecedents too. Further, our results show that particular anaphors seem to prefer certain types of antecedents: This (as well as the antecedent types themselves) can inform the design of feature sets for future non-NA anaphora resolution models.

In some ways, the results of our analysis conform to our intuitions about the behavior of particular anaphors: The shell noun *Tatsache* 'fact' seems to prefer surface forms thought to correspond to propositional abstract entities, whereas *Ziel* 'goal' and *Absicht* 'intent' strongly prefer infinitival VPs that are usually associated with actions.

We also see that the various anaphor lemmas—both pronouns as well as shell nouns—exhibit at times highly divergent properties, and this is a finding that has important implications for computational approaches, since future designs must allow for such divergences and treat novel anaphoric expressions accordingly.

Acknowledgments

We would like to thank our reviewers for their detailed and helpful comments and suggestions.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on language and computation*, 2(4):597–620.

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

Donna K. Byron. 2004. *Resolving pronominal reference to abstract entities*. Ph.D. thesis, University of Rochester, Rochester, NY, USA.

Alena Böhmová, Jan Hajič, Eva Hajičcoá, and Barbora Hladká. 2003. The Prague Dependency Treebank. In Anne Abeillé, editor, *Treebanks: building and using parsed corpora*, pages 103–127. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Manfred Consten, Mareile Knees, and Monika Schwarz-Friesel. 2007. The function of complex anaphors in texts. Evidence from corpus studies and ontological considerations. In Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees, editors, *Anaphors in Text: Cognitive, formal and applied approaches to anaphoric reference*, pages 81–102. John Benjamins, Amsterdam / Philadelphia.

Stefanie Dipper and Heike Zinsmeister. 2012. Annotating abstract anaphora. *Language Resources and Evaluation*, 46(1):37–52.

Stefanie Dipper, Christine Rieger, Melanie Seiss, and Heike Zinsmeister. 2011. Abstract anaphors in German and English. In *Anaphora Processing and Applications*, volume 7099 of *Lecture Notes in Computer Science*, pages 96–107, Berlin / Heidelberg, Germany. Springer.

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

Gill Francis. 1986. *Anaphoric Nouns*. Discourse Analysis Monographs 11, Birmingham: English Language Research, University of Birmingham.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference*, pages 57–60, New York City, NY, USA.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Varada Kolhatkar and Graeme Hirst. 2012. Resolving "this-issue" anaphora. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1255–1265, Jeju Island, Korea.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora with non-nominal antecedents in computational linguistics: A survey. *Computational Linguistics*, 44(3). Accepted for publication.

Klaus Krippendorff. 1995. On the reliability of unitizing contiguous data. *Sociological Methodology*, 25:47–76.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, third edition. See also: http://web.asc.upenn.edu/usr/krippendorff/m-Replacementofsection12.4onunitizingcontinuainCA,3rded.pdf.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Costanza Navarretta. 2007. A contrastive analysis of abstract anaphora in Danish, English and Italian. In *Proceedings of DAARC 2007 – 6th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 103–109, Lagos, Portugal.

Rebecca J. Passonneau. 1989. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 51–59, Vancouver, British Columbia, Canada.

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

Rebecca J. Schiffman. 1985. *Discourse constraints on 'it' and 'that': A study of language use in career-counseling interviews*. Ph.D. thesis, Faculty of the Division of the Humanities, Department of Linguistics, University of Chicago.

Hans-Jörg Schmid. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Topics in English Linguistics 34. Mouton de Gruyter, Berlin, Germany.

Fabian Simonjetz and Adam Roussel. 2016. Crosslinguistic annotation of German and English shell noun complexes. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, pages 265–278, Bochum, Germany.

Fabian Simonjetz. 2015. Retrieving German shell nouns using dependency patterns. http://www.researchgate.net/publication/306020586_Retrieving_German_Shell_Nouns_Using_Dependency_Patterns.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2018. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*. To appear.

Bonnie Lynn Webber. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122, Buffalo, NY, USA.

Heike Zinsmeister, Stefanie Dipper, and Melanie Seiss. 2012. Abstract pronominal anaphors and label nouns in German and English: selected case studies and quantitative investigations. *TC3. Translation: Computation, Corpora, Cognition*, 2(1):47–80.

A Appendix: Tables and Figures

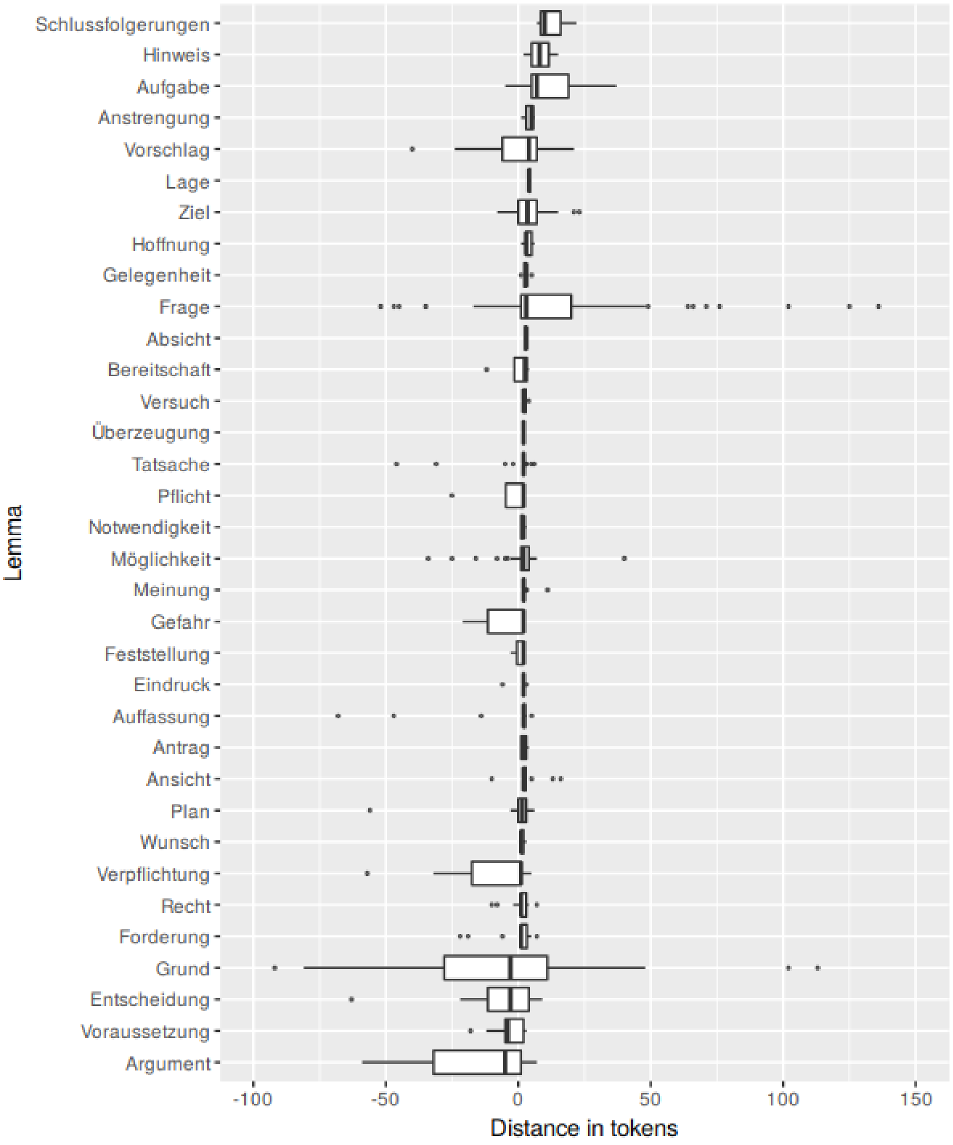

Figure 3: Distances between shell nouns and their antecedents. Negative distances indicate true anaphoric relations, i.e., the antecedents precede the anaphors, whereas positive distances indicate what are more accurately termed cataphoric instances.

	SENT	QUEST	DASS	WH-IND	ZU	V-FIN	V-PART	V-OTHER	NP	PP	OTHER
das	0.37	–	0.06	0.01	0.09	0.31	0.01	0.05	0.08	0.01	0.01
dies	0.29	–	0.27	–	0.13	0.2	–	0.07	0.04	–	–
es	0.21	–	0.09	–	0.12	0.21	0.06	0.09	0.21	–	–
Absicht	–	–	–	–	1	–	–	–	–	–	–
Ansicht	–	–	0.67	–	0.22	0.11	–	–	–	–	–
Anstrengung	–	–	–	–	0.67	–	–	–	–	–	0.33
Antrag	–	–	–	–	0.17	–	–	–	–	0.83	–
Argument	–	–	0.67	–	0.33	–	–	–	–	–	–
Auffassung	–	–	0.74	–	0.16	0.11	–	–	–	–	–
Aufgabe	–	–	–	–	0.38	–	–	–	0.46	0.15	–
Bereitschaft	–	–	–	–	0.75	–	–	0.25	–	–	–
Eindruck	–	–	0.43	–	–	0.57	–	–	–	–	–
Entscheidung	0.07	–	0.13	0.13	0.4	–	–	0.07	0.13	0.07	–
Feststellung	0.33	–	0.33	–	–	0.33	–	–	–	–	–
Forderung	–	–	0.2	–	0.2	–	–	0.05	0.05	0.5	–
Frage	0.01	0.43	0.04	0.2	0.01	–	–	–	0.2	0.09	–
Gefahr	–	–	0.43	–	0.14	–	0.14	–	–	0.29	–
Gelegenheit	–	–	–	–	0.83	–	–	–	–	–	0.17
Grund	0.68	–	0.08	–	0.04	–	–	–	0.12	0.08	–
Hinweis	–	–	1	–	–	–	–	–	–	–	–
Hoffnung	–	–	0.6	–	0.2	–	–	–	–	0.2	–
Lage	0.33	–	–	–	0.67	–	–	–	–	–	–
Meinung	–	–	0.83	–	0.08	0.08	–	–	–	–	–
Möglichkeit	–	–	0.15	–	0.62	–	–	–	0.15	0.06	0.03
Notwendigkeit	–	–	0.14	–	0.43	–	–	–	0.43	–	–
Pflicht	–	–	–	–	0.75	–	–	–	–	0.25	–
Plan	–	–	–	–	0.62	–	–	–	0.25	–	0.12
Recht	–	–	–	–	0.28	–	–	0.06	0.06	0.61	–
Schlussfolgerung	0.5	–	0.25	–	–	–	–	–	0.25	–	–
Tatsache	0.11	–	0.79	–	0.05	0.05	–	–	–	–	–
Überzeugung	–	–	0.67	–	0.33	–	–	–	–	–	–
Verpflichtung	–	–	–	–	0.29	–	–	–	0.14	–	0.57
Versuch	–	–	–	–	0.86	–	–	0.14	–	–	–
Voraussetzung	–	–	0.11	–	–	0.22	–	–	0.33	0.22	0.11
Vorschlag	0.05	–	0.05	–	0.43	–	–	–	0.33	0.14	–
Wunsch	–	–	–	–	0.33	–	–	–	0.33	0.33	–
Ziel	0.08	–	0.04	–	0.32	–	–	–	0.48	0.08	–
All pronouns	0.29	–	0.14	0.00	0.11	0.24	0.02	0.07	0.11	0.00	0.00
All shell nouns	0.06	0.01	0.25	0.01	0.34	0.04	0.00	0.02	0.11	0.11	0.04

Table 2: Per-lemma distribution of antecedent types.

Type	Condition on the antecedent
SENT	last token is "."
QUEST	last token is "?"
DASS	first lemma is *dass*/*daß* 'that'
WH-IND	first lemma is *ob* 'whether' or an interrogative pronoun (embedded/no final "?")
ZU	antecedent contains a token tagged as VVIZU or PTKZU
V-FIN	head is a finite verb (including modals and auxiliaries)
V-PART	head is a participle verb (including modals and auxiliaries)
V-OTHER	head is some other verb form (including modals and auxiliaries)
NP	antecedent is an NP
PP	antecedent is a PP
OTHER	antecedent is some other category

Table 3: Description of antecedent types.

Lemma	A_o	Lemma (cont'd)	A_o	Lemma (cont'd)	A_o
Annahme	1.00	Entscheidung	0.90	Forderung	0.85
Ansicht	1.00	Meinung	0.90	Ziel	0.85
Aufgabe	1.00	Bemühung	0.89	Antrag	0.83
Eindruck	1.00	Hoffnung	0.89	Plan	0.83
Hinweis	1.00	Recht	0.89	Standpunkt	0.83
Schlussfolgerung	1.00	Pflicht	0.88	Vorschlag	0.83
Tatsache	1.00	Umstand	0.87	Gefahr	0.77
Überzeugung	1.00	Grund	0.86	Frage	0.75
Wille	1.00	Lage	0.86	Verpflichtung	0.73
Möglichkeit	0.93	Notwendigkeit	0.86	Voraussetzung	0.73
Gelegenheit	0.92	Versuch	0.86	Schwierigkeit	0.71
Auffassung	0.91	Wunsch	0.86	Bedürfnis	0.60

Table 4: Agreement on the non-NA anaphoric status of shell noun candidates by lemma. (Only cases with $n > 5$ shown.)

Type	Divergent units	Pronouns	Shell nouns
SENT	sentence (including coordinated ones)	0.06	0.31
C-SUB	subordinate clause	0.02	0.12
CONJ	conjunction	0.08	0.02
V-FIN	finite main verb	0.02	0.00
M-FIN	finite modal verb	0.04	0.00
A-FIN	finite auxiliary	0.06	0.02
V-NFIN	non-finite main verb (participles, infinitives)	0.01	0.00
SUBJ	subject is missing or differs	0.14	0.05
ADV	adverbials, negation particles	0.05	0.02
C	matrix clause	0.05	0.07
PUNCT	punctuation	0.29	0.05
ALL	no overlap	0.11	0.26
UNDEC	too little context to label	0.01	0.02
NP	one annotator marked a nominal antecedent	0.03	0.00
OTHER		0.06	0.12

Table 5: Antecedent divergences by type.