Manual for the Annotation of in-document Referential Relations

Author: Karin Naumann (M.A.),
Seminar für Sprachwissenschaft, Abt. Computerlinguistik Universität Tübingen

Date: July 2006

Abstract

This paper presents relevant information concerning our annotation of in-document coreference and anaphora / cataphora. It provides a definition of the textual and semantic relation types and the category system used for the annotation together with a description of potential markables in the framework of coreference and anaphora/cataphora resolution. It also describes the data base containing the annotated texts and gives an illustration of the annotation tools used for our task. The overall aim is to provide comprehensive and comprehensible guidelines for both users of our released data and researchers designing a similar task. Therefore, it does not only describe the annotation background and process but also unfolds the process of discussing and deciding on controversial cases in order to arrive at a reliable annotation standard.[1]

Contents

1 Introduction 2
1.1 Relation Types 2
1.2 Annotation Data 2
1.3 Potential Markables 3
1.4 Category System 4

2 Discussion of MATE Categories 5

3 Annotation of Referential Relations 9
3.1 Identification and Extraction of Markables 9
3.2 Annotation Tool 9
3.3 Annotation Categories 11
3.3.1 Coreferential 11
3.3.2 Anaphoric 15
3.3.3 Cataphoric 17
3.3.4 Bound 18
3.3.5 Split Antecedent 20
3.3.6 Instance 21
3.3.7 Expletive 23

4 Conclusion 26

5 References 26

[1] Many thanks to Heike Zinsmeister, Holger Wunsch, Yannick Versley and Piklu Gupta for helping me refining content and form of the text.

1 Introduction

1.1 Relation Types

The linguistic relations that are annotated in the texts can be subsumed under the notion of "referential relations". We define referential relations as a cover-term for all contextually dependent reference relations. The annotation is restricted to anaphoric, cataphoric and coreferential expressions referring to a nominal or pronominal antecedent preceding or following the dependent expression within the same text. We follow Mitkov (2003) in defining anaphora as a linguistic phenomenon of pointing back to a textual entity preceding the anaphor, i.e. the referring phrase. In the case of cataphora, the antecedent which is referred to is following the cataphor. Thus, the linguistic framework underlying this definition is on the level of text rather than on the level of semantics. Coreference, on the other hand, relates to semantic-pragmatic reference of two or more nominal phrases to the same extra-linguistic referent. It follows from this definition that there might be cases of anaphora where the referring terms are not coreferential. These cases are called "identity-of-sense anaphora", not "identity-of-reference anaphora".

In our framework, we accomplish the task of coreference resolution, thus identifying all coreferential chains in a text. We also annotate all instances of identity-of-reference anaphora. With respect to identity-of-sense anaphora, we annotate all instances of so-called "bound" anaphora. But note that in the current release, these cases are not included in the data in order to get a maximum of annotation consistency, this category being rather problematic. Nevertheless, those instances have been annotated and will be published in a future release of our data.
We do not annotate other purely semantic relations, e.g. 'part-whole' or holonymy-metonymy. We also disregard the annotation of event anaphora, i.e. those relations which hold between an anaphoric expression and an underlying proposition or subsumed event rather than an identifiable antecedent in the text. Concerning the word class, we annotate instances of pronominal anaphora and coreferential lexical NP anaphora but neither verb and adverb anaphora nor zero anaphora in elliptic constructions.

The inventory of the relations is inspired by the annotation scheme first developed in the MATE project (cf. Davies et al.: 1998). Nonetheless, it only adopts those referential relations from MATE which correspond to the definition of relation types described above.

1.2 Annotation Data

The annotation is based on the Tübingen treebank of written German (TüBa-D/Z). This treebank uses as its data source a collection of articles of the German daily newspaper "taz" (i.e. "die tageszeitung"). From this database, currently 1102 data files, i.e. newspaper articles, with 23.496 sentences comprising 407.879 tokens are annotated for referential relations.

Due to its fine grained syntactic annotation, the TüBa-D/Z treebank data are ideally suited as a basis for the identification of markables and for extracting relevant syntactic and semantic information for each markable. For further information, please refer to the corresponding annotation manual (Telljohann et al.: 2006).

1.3 Potential Markables

The markables that are subject to annotation are those linguistic elements that might refer to a contextual antecedent. Poesio (2004) defines markables as “the text constituents that realize semantic objects that may enter in anaphoric relations”. Hirschman and Chinchor (1997) state “... that just because an element is "markable", it does not follow that there are later references to it -- that is, it may or may not participate in coreference. [...]. The relation is marked only between pairs of elements both of which are markables. This means that some markables that look anaphoric will not be coded, including pronouns, demonstratives, and definite NPs whose antecedent is a clause rather than a markable.“

Markables that will be annotated for anaphora/cataphora and coreference are definite noun phrases, personal pronouns, relative, reflexive, and reciprocal pronouns, demonstrative, indefinite and possessive pronouns. All markables are extracted automatically from the TüBa-D/Z treebank. With regard to the text type and text position, we annotate all parts of the newspaper article, i.e. we annotate the proper article text as well as all headlines, subtitles and lists, enumerations, etc. preceding/following the text or being inserted in the article.[2]

We annotate coreferential definite descriptions, i.e. definite NPs, including complex (e.g. coordinated) noun phrases. The boundaries of a nominal markable are defined by the maximal extension of this NP which includes complements as well as adjuncts of the head noun, cf. example 1.1.

Example 1.1: TüBa-D/Z, file 235

Eine niederländische Sozialhilfeempfängerin hat in einem alten Buch, das die Frau für umgerechnet zwei Mark auf einem Flohmarkt in Utrecht erstanden hatte, [1 zwei Rembrandt-Radierungen] gefunden. Zu Hause fielen ihr [2 die beiden Graphikblätter [[des Meisters] [aus dem 17. Jahrhundert]]] entgegen.

Engl[3].: A welfare recipient from the Netherlands found [two Rembrandt etchings] in an old book that the woman had bought at a flea market in Utrecht for two Deutschmark in local currency. At home, [both graphics of the 17th century master] fell into her hands.

The maximal NP also includes appositions and other statements in brackets, between commas and dashes. As mentioned before, we also establish relations between elements of the headline and the text itself. Example 1.2 and 1.3 illustrate this fact.

Example 1.2: TüBa-D/Z, file 104

[1 Stahmer]: 3.000 Schüler aus Bonn erwartet.

[...] Wie [2 Schulsenatorin Ingrid Stahmer (SPD)] sagte, seien bislang jedoch „nur einige hundert‟ angemeldet worden.

Engl.: [Stahmer]: 3.000 pupils from Bonn expected. [...] As [school senator Ingrid Stahmer (SPD)] said, ‟only a few hundred” had been registered up to now.

[2] Graphs, tables and illustrations are not included in the database.

[3] The English translation of all examples in this document aims at being able to understand the sense of the German equivalent but also at reflecting the German sentence structures in order to be able to follow the argumentation.

An NP markable might itself be embedded in another NP. In example 1.3, markable 1 “Radunskis” functions as prenominal modifier within the complex NP “Radunskis neue Entschlußfreude”.

Example 1.3: TüBa-D/Z, file 240

[[1 Radunskis] neue Entschlußfreude].

Kurz vor Ende der Wahlperiode dreht [2 der Senator] noch einmal voll auf.

Engl.: [Radunkski’s] new delight in deciding. Shortly before the end of the legislative period [the senator] gives his all.

MATE uses categories for possessive relations within a noun phrase. Example 1.4. illustrates the relation type “attribute”. In example 1.5, there exists a “part”-relation between markable 1 and 2. In contrast to the MATE annotation scheme (cf.: Poesio 2000), we do not annotate relations within a phrase. You will find a brief discussion on the differences between MATE and our annotation in section 2.

Example 1.4:

[1 The height] of [2 the wings]

Example 1.5:

[1 The seat] of [2 the chair]…

Regarding the annotation of referential relations, we do not annotate indefinite nominal descriptions containing an indefinite article or a quantifier. They may be antecedents, though. Thus, in example 1.6 we annotate the coreference relation between [1] and [4] but not the one between [2] and [3].

Example 1.6: TüBa-D/Z, file 303

[1 Ein Bundeswehrsoldat] ist am Dienstag in Griechenland von [2 einem Nato-Gegner] verletzt worden. Medienberichten zufolge schleuderte [3 ein 29 Jahre alter Grieche] in Salonica seinen Helm auf die Windschutzscheibe eines Militärfahrzeugs, das [4 der Soldat] fuhr.

Engl.: On Tuesday, [a soldier of the German Federal Armed Forces] was injured by [an opponent of Nato] in Greece. According to media reports, [a 29-year-old Greek] in Salonica hurled his helmet onto the windscreen of a military vehicle that [the soldier] was driving.

1.4 Category System

As mentioned above, the categories are based on the annotation scheme developed within the MATE project. It is, however, an adaption of this system, i.e. those categories in MATE not corresponding to the definition of markables and relation type described in section 1.1. and 1.2 are either avoided or redefined.

We use the following MATE categories for complex relations:

Our labels	Corresponding MATE labels
bound	bound anaphors
split_antecedent	element
split_antecedent	evoked entities
instance	instantiation

We do not use the following MATE categories: “function-value”, “subset”, “attribute”, “part”, “strict possession”, “event relation”, “situation” and “empty strings”. An explanation for the use or rejection of MATE categories can be found in section 2.

In conclusion, our category system currently comprises the following categories for the annotation of referential relations:

- “coreferential”
- “anaphoric”
- “cataphoric”
- “bound”
- “split_antecedent”
- “instance”
- “expletive”

For an extensive description of the categories please refer to section 3.3.

2 Discussion of the MATE Category System

Since the annotation of coreferential, anaphoric and cataphoric relations apparently causes few problems, the first step in developing an annotation standard was to discuss the MATE categories for the annotation of “complex relations”.

There are two questions involved in this task:

1.) Which MATE categories should be used according to our definition of referential relations and potential markables?
2.) Which labels should be used for identifying our own annotation categories?

In the following, the MATE categories for complex relations are defined and illustrated by examples from the MATE annotation manual (Poesio: 2000). The illustration is accompanied by a short explanation for our decision to use or to reject the respective category.

Bound Anaphors

The anaphor is bound by the same quantifier as its antecedent.

Example 2.1: `[1 Nobody] likes to lose [2 his] job`
Example 2.2: `[1 Every man] for [2 himself]`

The category is used under the same label “bound” because it agrees with the definition of anaphora used in our project and differs from the description of the category “anaphoric” because it does not includes a definite description as its antecedent and thus does not imply coreference.

Function-Value

Relationship between a function and its value(s).

Example 2.3:
`[1 The temperature] rose to [2 90 degrees] before dropping to [3 70 degrees]`

Describing an exclusively semantic relation, this category is not used in our project.

Set Relations

a) Element: one discourse entity is an element of the set denoted by the other discourse entity.

Example 2.4:
`[1 The kids] went to a party last weekend. [2 Paul] wanted to wear his new suit, but [3 Jane] insisted on wearing her jeans`

The category complies with our definition of anaphora and it is used with the label "split antecedent" (formerly "part_of"). For further discussion of this category see section 3.3.5.

b) Subset: one discourse entity denotes a subset of the set denoted by the other discourse entity.

Example 2.5:
`Alors donc vous avez ici les modèles de [1 fusées]. […] Et vous allez de vous mettre d'accord sur un classement/ hein classer [2 les fusées qui ont bien volé] ou [3 les fusées qui ont moins bien volé].`
`(Engl.: Here you have the models of the [1 rockets]. Please classify [2 the rockets which flew well], and [3 the rockets which didn't fly well].)`

The category denotes a semantic relation. Therefore, it is not used in our project.

c) Possessive Relations

Attribute: one discourse entity expresses something which is an attribute of another discourse entity.

Example 2.6: `[1 The height] of [2 the wings]`

On the one hand, this category denotes a semantic relation which does not correspond to our definition of referential relation. On the other hand, it does not correspond to the definition of markable boundaries because it is assigned to a relation within the nominal phrase.

Part: One discourse entity denotes a physical part of another discourse entity.

Example 2.7: [1 The seat] of [2 the chair]

The category denotes a semantic relation within a phrase. Therefore, it is not used in our project.

Strict possession: the relationship between two objects where one 'belongs' to the other.

Example 2.8:
`It was a brave decision by [1 Jerry Seinfeld] to turn down $5m an episode to make another series of [2 [3 his] hugely popular sitcom]`

It is not used in our project being a semantic relation. Note that the anaphoric relation between the possessive pronoun "his" (markable 3) and "Jerry Seinfeld" (markable 1) is annotated.

Instantiation

The second discourse entity refers to a particular instantiation of the class identified by the first discourse entity.

Example 2.9:
`Speaker A:"We need [1 oranges]." - Speaker B: "There are [2 some] at Corning"`

This category corresponds to the definition of the relation type "identity-of-sense anaphora" and is used under the label "instance" in the same sense.

Event Relation

Link between a discourse entity and a preceding event, expressed by a verbal phrase or sentence.

Example 2.10:
`[1 Muslims from all over the world were taught gun-making and guerrilla warfare in Afghanistan]. [2 The instructors] were members of some of the most radical Islamic militant groups in the region.`

This category does not denote a contextual relation but it refers to event anaphora resp. frame semantics. It is not used in our project.

Situation

Link between an item or situation, and typical components of that scenario.

Example 2.11:
`[1 The victorious French team] is parading [2 the World Cup] through [3 Paris] from the top of [4 a double-decker bus] as [5 a nation] continues with [6 mass celebrations].`

This category also belongs to the field of frame semantics. It does not denote a referential relation and is not used in our project.

Propositions, Events and Actions

Link between a discourse entity and a previous proposition, event or action.

Example 2.12:
[1 The 23-year-old had hit his head against another player] during a game of Aussie-rules football. McGlinn remembered nothing of [2 the collision], but developed a headache and had several seizures.

Denoting event anaphora rather than referential relations on text level, this category is not used in our project.

Evoked entities

These elements do not contain strings from the text or dialogue, but an informal description of the evoked entity.

Example 2.13:
[1 John] arrived at 7, but [2 Mary] was much later. [3 They] missed the film and went to the bar instead.

Being very close to the description of the set relation "element" (see above), we decided to annotate both cases under the label of "split_antecedent" (formerly "part_of").

Empty Strings and Clitics

For omitted elements that have to be reconstructed from the context.

Example 2.14:
Add [1 the dry yeast] to the water and let [2 *empty*] sit for a few minutes.

Example 2.15:
Speaker A: Dov'e [1 Gianni?] - Speaker B: [2 *empty*] È andato a mangiare.

The category is used for elliptic constructions. The relation described here is not a contextual relation but it includes bridging assumptions and presuppositions. Thus, it is not annotated in our project.

3 Annotation of Referential Relations

This section deals with the identification and extraction of markables. It gives a description of the annotation tool(s) used for the mark-up process and explains the annotation categories mentioned above.

3.1 Identification and Extraction of Markables

Annotation of referential relations involves two main tasks:

1.) the identification of markables, i.e. identifying the class of expressions that can enter into referential relations
2.) the identification of the particular referential relations that two or more expressions enter into.

Identification of markables requires at least partial syntactic annotation of the text. Markables have to be identified semi-automatically from the output of a chunker or full parser, if referential relations have to be annotated from plain text. If a parser is not available, the markables have to be identified completely manually. However, in each of these two scenarios, identification of markables is a time-consuming process. In case of semi-automatic annotation, the effort depends on the quality of the parser, but will require at least some amount of manual postcorrection of the parser output. Identification of markables is considerably easier for treebank data since treebanks already provide the necessary syntactic information. For German, there are currently two large-scale treebanks available: the NEGRA/TIGER treebank (Brants et al., 2002) and the Tübingen treebanks for spoken and written German (Stegmann et al.: 2000; Telljohann et al.: 2006). All the treebanks were annotated with the help of the annotation tool "Annotate" (Plaehn, 1998). The treebank annotations are available in three different formats: the NEGRA export format (Brants, 1997), an XML format[4] (cf. appendix of this stylebook), and the Penn treebank bracketing format[5] (Marcus et al.: 1993).

3.2 Annotation Tool(s)

The annotation of referential relations is performed manually. Until the beginning of 2006, the task was carried out by means of the annotation tool MMAX (Müller and Strube, 2003), developed by the European Media Lab (EML)[6]. The abbreviation "MMAX" stands for "Multi-Modal Annotation in XML". It is a stand-off annotation, i.e. base data and annotation data are stored in separate files. The annotation categories are considered as attribute values assigned to the respective markables.

MMAX recognizes two different relation types:

1.) **Set Relation**: transitive and undirected
2.) **Pointer Relation**: intransitive and directed

[4] also see: http://www.sfs.uni-tuebingen.de/en_exportxml.shtml, visited 20.06.2006
[5] also see: http://www.cis.upenn.edu/~treebank, visited 20.06.2006
[6] cf.: http://mmax.eml-research.de, visited 20.06.2006

In our annotation, the set relation was used for the annotation categories “anaphoric”, “cataphoric”, “coreferential” and “bound”. The pointer relation was used for the categories “split_antecedent” (formerly “part_of”) and “instance”.

The annotation of referential relations with MMAX included the following steps:

- **Step 1:** in the text window, the markable that is to be annotated is selected by left mouse click (highlighted with yellow background colour). Note: you do not have to create the markables. They are taken from the treebank and displayed in specific font colour.
- **Step 2:** in the attribute window, an attribute value at the level “Type” is chosen and assigned to the markable
- **Step 3:** in the text window, the relation between markable and antecedent is established by clicking on the immediate antecedent (except for cataphora) with right mouse click and choosing between set or pointer relation

=> Visualized result: all members of one set are highlighted (red font colour).

Since February 2006, we use the Annotation Tool “PALinkA”[7] developed by Constantin Orasan from Wolverhampton University[8]. It is a java-based application which uses XML format for the annotation data. Beside the visualization of referential chains and different search functions, the tool provides the possibility to add comments to each annotation which is very helpful for revising and discussing problematic cases.

The annotation with PALinkA is carried out in the following manner:

- **Step 1:** select menu “Tags” from the menu bar at the top of the browser window and choose the annotation category that you want to apply. Note: You do not have to create the markables. This information is already available from the treebank and is displayed by square brackets plus background colour. Additionally, all markables are displayed in alphabetical order in the right window frame.
- **Step 2:** click on the markable that you want to annotate. Note: if the markable is itself part of a larger markable, you have to select the item from a menu window.
- **Step 3:** click on the closest antecedent (in the case of cataphora, it is the immediately following NP)
- **Step 4**: add comment and save the annotation

=> **Visualized result:** the referential expressions receive the same background colour. When moving the computer mouse over the anaphor, an arrow will appear pointing to the antecedent.

[7] See: http://clg.wlv.ac.uk/projects/PALinkA/, visited 20.06.2006

[8] See: http://clg.wlv.ac.uk/, visited 20.06.2006

3.3 Annotation Categories

In this section, the categories for the annotation of referential relations are defined and illustrated by examples from the text corpus. Additionally, the section quotes problematic cases connected with the development of an annotation standard for each annotation category.

3.3.1 Coreferential

This category is used for definite NPs in cases where two NPs refer to the same extra-linguistic referent. This definition follows van Deemter and Kibble (2000).

Example 3.1: TüBa-D/Z, file 238

```
Der Vorhang geht wieder auf im [1 Metropol]. Kultursenator will [2 das Theater] an Privatinvestor verkaufen.

Engl.: The curtain rises again in [the Metropol]. Culture senator wants to sell [the theatre] to private investor.
```

Note that often there are semantic relations involved that hold between two coreferential expressions. In this example, it is the relation of "instance/class". In a forthcoming project phase, this kind of information may be used to identify potential markables and their antecedents which could then be extracted automatically from GermaNet, a lexical-semantic net for German (Kunze/Lemnitzer: 2002).

Discussion / Problematic cases

When defining the category system for the annotation of coreference, we decided not to consider semantic relations between textual entities nor structural characteristics, e.g. identical head of two or more NPs. Instead, we restricted the definition of coreference to a purely functional, pragmatic description, i.e. reference to the same extra-linguistic referent.

The following section exhibits some problematic cases regarding coreferential relations that we encountered during the annotation of the TüBa-D/Z treebank.

For appositions and copula constructions we made the following decisions:
Appositions belonging to the same maximal NP -> no internal reference marking

- „Gerhard Schröder, ein passionierter Zigarrenraucher,..." -> one NP[9]
  Engl.: „Gerhard Schröder, a passionate smoker of cigars,..."
- „Gerhard Schröder, (der) Bundeskanzler, ..." -> one NP
  Engl.: „Gerhard Schröder, (the) Federal Chancellor,..."

Copula Constructions:

- „Gerhard Schröder ist der deutsche Bundeskanzler" -> definite predicate NP -> coreferential NPs.
  Engl.: Gerhard Schröder is the German Federal Chancellor
- „Gerhard Schröder ist ein passionierter Zigarrenraucher" -> indefinite predicate NP -> not coreferential
  Engl.: Gerhard Schröder is a passionate smoker of cigars.

[9] All „invented" examples that are not taken from the Treebank appear without number and with normal font type

Constructions with „als“ (engl. “as”):
Example 3.2 illustrates that NPs with „als“ (engl. “as”) do not enter into a coreferential relation.

Example 3.2: TüBa-D/Z, file10
Es gibt [1 einen neuen Kuli] auf dem Markt , der heute schon als [2 Rarität] zu bezeichnen ist.

Engl.: There is [a new ball-pen] on the market, which today is already considered [a rarity].

We agreed that predicative uses of NPs should generally be treated separately. Nevertheless, the definition of a new annotation type is postponed to a further project phase.

Another problem concerns the definiteness of markables: should we also annotate indefinite descriptions sharing a coreference relation with a textual antecedent being indefinite because of the text type? We decided not to do so (cf. section 1.3, example 1.6).

As for generic uses of NPs, we decided not to create coreferential relations. But concerning abstract concepts, we do annotate the nominal expression, if it appears repeatedly in the text with the same interpretation.

Another problem concerns an NP referring to a person which acts in different roles. In example 3.3, we have to differentiate between the actor „John Travolta“ (markable 1) and his role as a “berechnender Karriereanwalt” (markable 3). In this case, we decided to put markables 1-5 into one set because in the given context all terms refer to the same person.

Example 3.3: TüBa-D/Z, file 180
[1 John Travolta] verklagt als Bostoner Anwalt zwei Firmen, die [2 er] für den Leukämietod von acht Kindern verantwortlich macht. Anfangs wittert [3 der berechnende Karriereanwalt] nur die hohe Entschädigungssumme, doch ganz allmählich wird der Fall zu einer selbstzerstörerischen Obsession. Gerichtsdrama, Umweltthriller und großes Schauspielkino, in dem [4 Travolta] und [5 sein] Gegenspieler Robert Duvall zu Hochform auflaufen.

Engl.: [John Travolta] as a lawyer from Boston sues two companies which [he] considers responsible for the death of eight children as a result of leukaemia. In the beginning, [the calculating high flying advocate] only scents high compensation sums, but slowly the case becomes a self-destroying obsession. Court drama, environmental thriller and great actor’s cinema, where [Travolta] and [his] antagonist Robert Duvall achieve top form.

The next problem is related to lacking gender/number/case agreement, which is especially important in German with its overt inflectional system. In example 3.4, markable 1 referring to a country and markable 2 referring to its inhabitants are coreferent despite the difference in number of the nominal head. Number, gender and case are not relevant regarding coreferential relations as long as the terms refer to the same referent, in this example the Finnish Icehockey Team.

Example 3.4: TüBa-D/Z, file 229
Schweden und [1 Finnland] im Viertelfinale der Eishockey-WM : Das 6:1 über die Schweiz war der zweite Sieg des Titelverteidigers , der damit genauso 4:0 Punkte hat wie [2 die Finnen] (4:1 über Weißrußland).

Engl.: Sweden and [Finland] in the quarter final of the ice-hockey world cup: the 6:1 score against Switzerland was the second victory of the title-holder, which has 4:0 points just like [the Finns] (4:1 over Belarus).

As an example for difference in case see example 3.5:

Example 3.5: TüBa-D/Z, file 202
Als Vorsitzender des HDO-Untersuchungsausschusses, der Mauscheleien von [1 SPD-Ministerpräsident Wolfgang Clement] bei der Oberhausener Trickfilmfirma aufklären soll, hat es Meyer vom wirtschaftspolitischen Sprecher binnen Halbjahresfrist zum Fraktionsvorsitzenden der NRW-CDU gebracht. Ohne die Pleitenserie [2 Clements] in den letzten Monaten wäre der Karriereschub des Duos nicht möglich gewesen.

Engl.: As chairman of the HDO commission of inquiry, which is to investigate irregularities by [SPD prime minister Wolfgang Clement] at the Oberhausen animation company, Meyer made it from the spokesman for economic policy to the faction leader of the NRW-CDU within the mid-year period. Without [Clement's] series of mishaps within the last months, the advance in the career of the two would not have been possible.

The next case illustrates the difference between identity of two strings on text surface and coreference between these strings. In example 3.6, there is no coreference relation between markable 1 and 2 because of the first markable referring to the person's name as a string of letters, not to the person itself.

Example 3.6: TüBa-D/Z, file 236
Was der Krieg weiterhin verspricht , ist viertens Abenteuer unter Lebensgefahr . Kann man aber schneller haben , wenn man sich in Amerika eine Zigarette ansteckt . Und alles unter diesem Level gibt 's schon : Bungee-Jumping , Wildwasser-Rafting und alle die Sachen , die hinten mit -ing aufhören wie [1 Sharping] ; [...]Frauen kriegen im zweitschlimmsten Fall ein Telegramm von [2 Sharping] und im schlimmsten Fall müssen sie sich die nächsten 30 Jahre Geschichten anhören , die anfangen mit " Damals im Kosovo ".

Engl.: What the war promises is fourthly adventure at the risk of one's life. One can achieve this faster by lighting up a cigarette in America. And everything beneath this level is already available: bungee jumping, wild water rafting and all those things ending in „-ing" like [Scharping]; [...] In the second worst case, women receive a telegram from [Scharping] and in the worst case, within the next 30 years they have to listen to stories beginning with „Once upon a time in Kosovo".

In example 3.7, we consider the title of a book mentioned in the text and the book itself as being coreferential.

Example 3.7: TüBa-D/Z, file 266
Seit er 1995 [1 die Textsammlung " Kanak Sprak "] veröffentlicht und damit deutsch-türkische Lebensentwürfe jenseits von multikultureller Idylle und Kulturkampfthesen vorgestellt hat , gilt er als Sprachrohr der dritten Einwanderungsgeneration : Feridun Zaimoglu , 1964 im türkischen Bolu geboren , seit gut 30 Jahren in Deutschland ansässig. [...]In [2 " Kanak Sprak "] schreiben Sie , die " Kanaks " seien " unerreichbar ".

Engl.: Since publishing [the collection of texts „Kanak Sprak"] in 1995 and thus presenting German-Turkish concepts of living beyond a multi-cultural idyll and theses of cultural conflicts, he is considered the spokesman of the third generation of immigrants: Feridun Zaimoglu, born in 1964 in the Turkish city of Bolu, living in Germany for almost 30 years. [...]. In ["Kanak Sprak"] you write that the "Kanaks" are "unreachable".

In example 3.8, the term „Opa“ (engl. grandfather) does not function as a term for family relations but as a proper name thus entering into referential relation with coreferential markable 2 and anaphoric pronoun (markable 3).

Example 3.8: TüBa-D/Z, file 236

So war [1 Opa] körperlich als auch mental gut auf den 14-18er vorbereitet . Auf den 39-45er war [2 Opa] als Ortsgruppenleiter ebenfalls bestens vorbereitet , durfte aber nicht mit , weil [3 er] schon zu alt war .

Engl.: In this way, [grandpa] was physically and mentally well prepared for the 14-18 (World War I). For the 39-45 (Word War II), [grandpa] as leader of a local group was likewise best prepared but was not allowed to join it because [he] was already too old.

The next problem concerns nominal coordination. In example 3.9, there exist coreferential relations between the coordinate NPs (markable 1 and 2) as well as between part of the NP (markable 3 and 4).

Example 3.9: TüBa-D/Z, file 103

[1 Grüne und ÖTV] kritisieren Privatisierungen
[2 [3 Die Grünen] und die ÖTV] haben die Privatisierungspolitik des CDU/SPD-Senats scharf kritisiert . Diese habe zu einer Verunsicherung der Beschäftigten geführt und eine Blockade für notwendige Modernisierungsmaßnahmen ausgelöst , hieß es gestern in einer gemeinsamen Presseerklärung der ÖTV-Chefin Susanne Stumpenhusen und der Fraktionsvorsitzenden von [4 Bündnis 90 / Die Grünen] , Michaele Schreyer.

Engl.: [Green Party and ÖTV] criticise privatisation. [[The Greens] and the ÖTV] strongly criticised the politics of privatisation by the CDU/SPD senate. It had led to uncertainty of the employees and caused a blockade against necessary means of modernisation, it was said yesterday in a joint press release of the ÖTV leader Susanne Stumpenhusen and the fraction leader of [Bündnis 90 / the Greens], Michaele Schreyer.

In example 3.10, there is no coreference relation between markable 2 and 3 because 2 functions as a label noun embedded in the definite expression “Staatsanwaltschaft Freiburg” and is not a referential expression.

Example 3.10: TüBa-D/Z, file 70

Aufgrund einer Anzeige des Bund für Umwelt und Naturschutz Deutschlands (BUND) ermittelt nun [1 die Staatsanwaltschaft [2 Freiburg]] . Der Verdacht , daß es sich um nicht zugelassene Maissorten handelt , wurde von der Chemischen Landesuntersuchungsanstalt in [3 Freiburg] bestätigt.

Engl.: Because of a complaint by the German Federation for Environment and Nature Conservancy (BUND) [the Public Prosecutor’s Department Freiburg] is investigating now. The suspicion that it concerns prohibited sorts of maize has been confirmed by the federal centre for chemical investigation in [Freiburg].

3.3.2 Anaphoric

This attribute value is used for definite pronouns referring back to a contextual antecedent. A set relation is established between the pronoun and its antecedent. In example 3.11 the NP is referred to by two reflexives (markables 2 and 5) and two personal pronouns (markables 3 and 4). Thus, markables 1-5 belong to the same set.

Example 3.11: TüBa-D/Z, file 336

[1 Ein klarer Ton] breitet [2 sich] aus, warm und satt, bis [3 er] den ganzen Saal erfüllt. Dann dünnt [4 er] aus, zerbröselt und verflüchtigt [5 sich].

Engl.: [A clear sound] spreads [itself] out, warm and full, until [it] fills the whole hall. After that, [it] thins out, falls into pieces and fades [itself].

Discussion / Problematic cases

There was a discussion regarding the annotation of reflexive pronouns because in German there are many cases of „inherently reflexive verbs“, as e.g. „sich ereignen“ (engl. „happen (itself)“). In these cases, the reflexive is not annotated because it does not refer. There are various tests for deciding whether a verb is inherently reflexive, e.g. coordination tests („Sie wäscht sich und ihre Schwester“, engl.: „she washes herself and her sister“ -> not inherently reflexive). These tests work out all right for clear-cut cases. But there are lots of boundary cases where this decision can not easily be made. Therefore, we decided to insert a comment for all reflexives that are not annotated for belonging to an inherently reflexive verb. Thus, it is possible to check all cases in a revisional phase after the annotation process. Additionally, the annotator refers to a list of verbs included in the IMSLEX lexicon (cf. Eckle-Kohler, 1999 and Fitschen, 2004) for deciding on these cases.

The next problem concerns personal pronouns. First person plural pronoun „wir“ (engl. “we”) in direct speech cannot easily be related to a speaker because it may include individuals which are not mentioned in the text. It is ambiguous, who the speaker includes in this group of people. Therefore, in example 3.12 we only establish an anaphoric relation between markable 3 and 4. But note that uses of “pluralis majestatis” are excluded from this rule, because they clearly refer to the speaker of the direct speech thus receiving the value “anaphoric” or “cataphoric”.

Example 3.12: TüBa-D/Z, file 149

[1 Izet] deutet auf die Plastiksandalen , die [2 ihm] geblieben sind. "[3 Wir] haben nur noch das , was [4 wir] auf dem Leibe tragen."

Engl.: [Izet] points to the plastic sandals, that have stayed with [him]. „[We] only have the things, that [we] are wearing on our bodies.‟

The next case is closely related to the last example. In example 3.13, there exists an anaphoric relation between markable 1 and 3, not 1 and 2, although the NP “meine Wenigkeit” stands for “me”. Remember: we establish anaphoric relations between elements of the text surface (in this example first person personal pronoun and first person possessive pronoun) not semantic relations.

Example 3.13: TüBa-D/Z, file 266

[1 Ich] sehe das eher als Plattform , die es möglich macht , daß Imran und [2 [3 meine] Wenigkeit] im Stakkato literarische Texte raushauen , Musikdarbietung inbegriffen.

Engl.: [I] consider this rather a forum that makes it possible that Imran and [[my] humble self] churn out literary texts in staccato, including musical performances.

The next topic concerns items closely related to personal and reflexive pronouns. In the TüBa-D/Z treebank, the intensifier „selbst“ (engl. “self”) is analyzed as an adverb and consequently does not belong to the group of potential markables. Therefore, it does not enter into an anaphoric relation. In example 3.14, only markable 3 is annotated referring anaphorically to markable 1.

Example 3.14: TüBa-D/Z, file 120

„Sollen [1 sie] mich doch bloß als Narren ansehen. Werden sehen, ob [2 [3 sie] selbst] nicht welche sind.‟

Engl.: „Let [them] consider me a fool. (They) will see, if [they [themselves] aren’t ones.‟

The next problem is related to the field of fixed expressions. The following two examples 3.15 and 3.16 exemplify idiomatic uses of pronouns. In these cases, the possessive pronouns are not annotated for anaphora since they do not refer to other lexical items within the text. Also compare inherently reflexive verbs (see above) and fixed expressions as e.g. “sich fragen”, “mit sich bringen”, “sich zeigen”, etc.

Example 3.15: TüBa-D/Z, file 033

„[Mein] Gott, müssen wir viele Kinder kriegen‟, sorgt sich ein frisch gebackener Bräutigam.

Engl.: [My] God, how many children we must have‟, the newly wed groom remarks sorrowfully.

Example 3.16, TüBa-D/Z, file 251

Und so pinseln und kleben, knipsen, schweißen und knobeln jetzt die Künstler, daß es nur so [seine] Art hat...

Engl.: And so the artists brush and paste, cut off, weld and toss, that it only has [it’s] way[10]...

With regard to interrogative pronouns, we only annotate those cases where these elements function as a relative pronoun, as e.g. in : „Ich interessiere mich für [1 das], [2 was] du sagst“ - engl.: I am interested in [that] [which] you are saying. In this case, markable 2 refers anaphorically to markable 1.

[10] The German idiomatic expression „seine Art haben“ means „it is a delight“

3.3.3 Cataphoric

A cataphoric relation holds between a pronoun referring to a following antecedent within the same or the following sentence. In contrast to the category “anaphoric”, the syntactic hierarchy overrules contextual criteria. Consequently, the attribute value is assigned to a definite pronoun preceding the antecedent of the main clause even if the extra-linguistic referent has already been mentioned within the preceding text.

In example 3.17, the personal pronouns „mich“ (markable 1) and „er“ (markable 3) receive the attribute value „cataphoric“, in the first case referring to the proper name “Ulrich Görlitz” (markable 2) and in the second case to the personal pronoun “er” (markable 5). On the other hand, markable 4 and markable 5 are annotated as “anaphoric” both referring to markable 2.

Example 3.17: TüBa-D/Z, file 120

“Sollen sie [1 mich] als Narren ansehen!” [2 Ulrich Görlitz] diente freiwillig in Hitlers Armee.[...] Aber [3 er] habe [4 sich] eben 1944 zur Armee gemeldet, um der Einberufung durch die Waffen-SS zu entgehen, erzählt [5 er].

Engl.: „Let them consider [me] a fool!“ [Urlich Görlitz] served in Hitler’s army voluntarily. [...] But [he] had enlisted [himself] in the army in 1944 in order to escape call up into the „Waffen-SS“, [he] explains.

Discussion / Problematic cases

Generally, the concept of cataphora can be defined in two different ways:
there exits a cataphoric relation between a referential expression and its contextual antecedent, if

1.) the antecedent has not been mentioned in the text yet (strict definition based on discourse level)
2.) the antecedent follows the referential expression within the sentence (weak definition based on sentence level)

We decided in favour of definition 2 in order to get the maximum of referential information. In a further working phase, it is still possible to sort out those cases that do not conform to definition 1.

In example 3.18, markable 2 receives the attribute value „cataphoric“ referring to the following antecedent markable 3 although the NP has already been mentioned in the heading of the article (markable 1). Relations between elements of the heading and elements of the article transgressing the boundaries of textual parts are predominated by internal relations within one textual part.

Example 3.18: TüBa-D/Z, file 102

Nach 222 Tagen Mahnwache am Alexanderplatz zogen [1 die Traktorenwerker aus Schönebeck] gestern ab. Ein Investor aus Nordrhein-Westfalen will 190 Arbeitsplätze sichern

222 Tage währte [2 ihre] kleine Traktoren-Mahnwache am Alexanderplatz. Mit drei landwirtschaftlichen Nutzfahrzeugen und ständiger personeller Präsenz im nebenstehenden Campingmobil dauerprotestierten [3 die Beschäftigten der Landtechnik Schönebeck (LTS)] sowie MitarbeiterInnen des Tochterunternehmens GS Fahrzeug- und Systemtechnik für die Übernahme ihres Betriebes durch einen West-Investor.

Engl.: After 222 days of solemn vigil at Alexanderplatz, [the workers on tractors from Schönebeck] withdrew yesterday. An investor from „Nordrhein-Westfalen‟ wants to save 190 jobs. [Their] little solemn vigil of tractors at „Alexanderplatz‟ lasted 222 days. With three agricultural commercial vehicles and a permanent personal presence in the neighbouring camper [the employees of the „Landtechnik Schönebeck (LTS)‟] as well as employees from the subsidiary „GS Fahrzeug- und Systemtechnik‟ were protesting permanently for the takeover of their company by a Western investor.

Note that further anaphoric expressions without a following nominal antecedent within the same sentence receive the attribute "anaphoric" as in example 3.19 where markable 3 refers back to markable 2.

Example 3.19: TüBa-D/Z, file 111
Daß [1 sie] von dieser Klausel Gebrauch machen würden , daran ließen [2 die eingeflogenen Damen und Herren] von Anfang an keinen Zweifel . Nur mit Mühe ertrugen [3 sie] den Versuch eines studentischen Personalratsvertreters , einen Mißbilligungsantrag gegen den FU-Kanzler vorzubringen.

Engl.: That [they] would use this clause, [the ladies and gentleman flown in] left no doubt from the beginning on. Only with effort did [they] suffer the attempt of a student representative of the personnel board to introduce a vote of no confidence against the chancellor of the „FU‟.

3.3.4 Bound

This category is used for set relations holding between a definite pronoun and a quantified or indefinite noun phrase / pronoun as its antecedent. The anaphor "is bound by the same quantifier as its antecedent" (Poesio 2000) (e.g. "some" or "many", etc.). As mentioned in section 1.1, this category is applied to all cases described here but is not included in the current release.

In example 3.20, this relation holds between markable 2 and the indefinite pronoun (markable 1).

Example 3.20: TüBa-D/Z, file 124
[1 Wer] einen Sitzplatz haben will, [2 der] muß um 19 Uhr, wer im Treppenhaus noch etwas hören will, sollte spätestens um 20 Uhr da sein.

Engl.: [Whoever] wants to get a seat, [he] has to be there at 7 p.m, whoever wants to hear something from the stairway should be there at 8 p.m. at the latest.

In case that there are further pronouns following the indefinite antecedent, these markables receive the attribute „anaphoric“, as shown in example 3.21. Markable 2 refers to the indefinite NP "Die meisten Benutzer" (markable 1) receiving the value "bound", whereas markable 3 receives the value "anaphoric" referring back to markable 2.

Example 3.21: TüBa-D/Z, file 94
[1 Die meisten Benutzer] kaufen [2 sich] [3 ihre] Tassen selbst.

Engl.: [Most of the users] buy [their] cups [themselves].

Discussion / Problematic Cases

One question concerning this category was how to treat generic uses of NPs (e.g. “der Deutsche an sich, der…”; “ein Mensch, der…”). We decided to restrict the use of the category “bound” to quantified non-referential antecedents but to include the impersonal 3rd person singular pronoun “man”.

It is important to differentiate between indefinite NPs including a quantifier which refer to an unspecified class of objects / persons and indefinite NPs referring to a specific group.

- [1 Einige Frauen] qualifizieren [2 sich] über spezielle Frauenförderungsmaßnahmen weiter. - Engl.: [Some women qualify [themselves] via special support schemes for women.
- [1 Einige Frauen] kamen auf mich zu, [2 die] mir unzählige Fragen stellten. - Engl.: [Some women] came to me [who] asked innumerable questions.

In the first example, markable 2 gets the attribute value “bound” referring to an unspecific indefinite group whereas in the second example it receives the value “anaphoric” referring to an indefinite but specific group of people.

Not all NPs including quantifiers are non-referential. Example 3.22 includes a quantified NP which refers to a specific group of people. Markable 2 therefore receives the value “anaphoric”, not “bound”.

Example 3.22, TüBa-D/Z, file 489

Der Selfmade-Diplomat kehrte mit [1 den drei in Jugoslawien inhaftierten Soldaten] aus Belgrad zurück.[...] [2 Sie] waren Mitglieder der dort stationierten UN-Friedenstruppen.

The self-made diplomat returned from Belgrade with [the three soldiers arrested in Yugoslavia]. [...] [They] were members of the UN peace corps based there.

As example 3.23 illustrates, lacking agreement (difference in number for markable 1 and 2) does not impede the use of the category “bound” (cf. example 3.4).

Example 3.23, TüBa-D/Z, file 172

„Wir werden doch [1 kein altes Hafenbecken] zuschütten“, sagte er, [2 die] seien doch der Reiz der geplanten neuen Hafen-City.

Engl.: „We will fill up [no ancient port basin]“, he said, [they] are the attraction of the planned new port city.

For our future work, we will create a list of text indicators including all possible surface triggers for this category.

3.3.5 Split_antecedent (formerly “part_of”)

The split_antecedent relation holds between coordinate NPs (e.g. “Jane and Mary”) or plural pronouns (e.g. “both”) and pronouns / definite NPs referring to one member of the plural expression. The category is not applied to semantic relations of the type “possessive relations”. In example 3.24, the indefinite pronoun “Sie” (markable 3) refers to both, the NP “eine junge Frau” (markable 1) and the coordinate NP “ihre gebrechliche Mutter und vier Kinder” (markable 2). A split_antecedent relation is established pointing from markable 3 to both of the antecedents.

Example 3.24: TüBa-D/Z, file 634

“Vor allem die letzten Stunden waren fürchterlich”, sagt [1 eine junge Frau], die [2 ihre gebrechliche Mutter und vier Kinder] über die Grenze führt. [3 Sie] sind zu Fuß gekommen, denn das Auto wurde [4 ihnen] von serbischen Freischärlern abgenommen.

Engl.: „Above all the last hours were terrible, said [a young woman] who is leading [her invalid mother and four children] across the border. [They] have come on foot because the car was taken from [them] by Serbian irregulars.

Note that there is an anaphoric relation between markable 4 “ihnen” and markable 3 “Sie”.

Discussion / Problematic cases

In the MATE project, possessive relations in the sense of semantic relations are annotated (see section 2). We decided, not to annotate this kind of relation but to introduce the category “split_antecedent” for reference to plural pronouns in the sense of the MATE category “evoked entities” and the set relation category “element”.

In the beginning of the project phase, the category was used under the label “part-of” (cf. Hinrichs et al.: 2004, 2005). The antecedents received the attribute value “part_of” and a pointer relation to the plural NP was established. A further step after discussing this category again was to reverse the relation, i.e. the plural expression receives the value “split_antecedent” pointing to the two or more antecedents.

In example 3.25, markable 3 receives the attribute value “split_antecedent” and points to both markable 1 and 2.

Example 3.25, TüBa-D/Z, file

taz: „[1 Sabrije], hast du noch Kontakt zu [2 deiner serbischen Freundin]‟? Sabrije: „Seit den Nato-Angriffen haben [3 wir] nicht mehr miteinander gesprochen.‟

Engl.: taz: „[Sabrije], are you still in contact with [your Serbian friend]‟? Sabrije: „Since the NATO attacks, [we] haven’t spoken to each other.‟

In example 3.26, markable 3 is not „part-of“ markable 1, but enters into a coreferential relation with markable 2.

Example 3.26, TüBa-D/Z, file 217

Dort lebt asketisch ein alter Zen-Meister mit [1 seinen beiden Schülern - Kibong, einem jungen Mönch, und [2 Haejin, einem Kind]]. [...] manchmal regnet es; oder der Milchzahn [3 des kleinen Jungen] wackelt.

Engl.: An old Zen master lives there ascetically with [his two disciples - Kibong, a young monk, and [Haejin, a child]]. [...] sometimes it rains; or the milk tooth of [the young boy] wobbles.

In example 3.27, markable 3 has split antecedents and therefore points to markable 1 and 2. If there was a coordinate NP [der All Jazz Club und das Villa], we could establish an anaphoric relation between markable 3 and the complex NP.

Example 3.27, TüBa-D/Z, file216

Nach Medienberichten müssen Ende Mai wegen neuer Eigentümer auch [1 der All Jazz Club] und Ende Juni [2 das Villa] schließen. [3 Beide] sind seit den fünfziger Jahren für ihre prominenten Gast-Konzerte bekannt.

Engl.: According to the news, at the end of May [the All Jazz Club] and at the end of June [the Villa] have to close because of new owners. Since the fifties, [both] have been well-known for their prominent concerts.

3.3.6 Instance

In cases where a specific pronoun or NP refers to a particular instantiation of the class identified by an NP, the category “instance” is assigned to the preceding or following pronoun / NP referring to that nominal antecedent. Thus, the specific expression is considered a specific instantiation of a class of entities. In example 3.28, the pronoun “jene” (markable 2) receives the value “instance” and points to the NP “viele Banken” (markable 1).

Example 3.28: TüBa-D/Z, file 414

Doch [1 viele Banken], vor allem [2 jene] im Staatsbesitz, sitzen auf Haufen von faulen Krediten.

Engl.: But [many banks], above all [those] belonging to the state, are sitting on a heap of rotten credits.

Discussion / Problematic Cases

In example 3.29, there is a purely semantic relation between markable 1 and 2. Thus, markable 2 is not annotated as instance of markable 1.

Example 3.29, TüBa-D/Z, file 239

Kurz vor Ende der revolutionären Demonstration knüppelte [1 die Polizei] in die Menge, während Steine und Glasflaschen in Richtung [2 der BeamtInnen] geworfen wurden.

Engl.: Shortly before the end of the revolutionary demonstration, [the police] were clubbing the crowd while stones and glass bottles were thrown in the direction of [the officers].

The following list represents all cases annotated so far. As you will notice, there are not very many. The list also includes those cases that were rejected after discussion of each sentence. Nevertheless, they are included in the list in order to illuminate the process of developing our standard. A corresponding remark will accompany these examples.

Example 3.30, TüBa-D/Z, file 046

„Jungen Linienrichtern wird bei [1 zweifelhaften Entscheidungen] - also **[2 solchen]** , [3 die] gegen BU ausfallen - schon mal angedroht , sie ins Internat zurückzuschicken .“
Engl.: „For [1 dubious decisions] - namely [2 those] [3 which] go against BU - young linesmen are threatened to be sent back to boarding school.

- ⇨ markable 2 = instance; markable 3 = bound

Example 3.31, TüBa-D/Z, file 055

„Die russische Diplomatie ist unentbehrlich , wenn es darum geht , [1 die zu erwartenden Störmanöver Miloevic'] zum Scheitern zu bringen . **[2 Solche Manöver]** sind todsicher zu erwarten , denn die G-8-Erklärung enthält keinerlei Details und keine Fristen , die den Rückzug der " militärischen , polizeilichen und paramilitärischen Kräfte aus dem Kosovo " regeln würden .“
Engl.: „The Russian diplomacy is indispensable when [1 the expected disruptive actions of Milosevic] have to be impeded. [2 Such actions] are dead sure because the G8 declaration contains no details and deadlines that regulate the fallback of the „military, police and paramilitary forces from Kosovo“.“

- ⇨ markable 2 = instance

Example 3.32, TüBa-D/Z, file 230

„Die konservativen Kräfte warten ja nur darauf , ihm [1 Sätze] um die Ohren zu hauen wie **[2 jenen]** von den 16 Mittelstrecklern , denen er in vier Wochen die Viererkette beibringe .“
Engl.: „The conservative powers are just waiting to bombard him with [1 sentences] like [2 the one] about the 16 middle-distance runners who he is teaching the double full-back formation within four weeks.“

- ⇨ markable 2 = instance

Example 3.33, TüBa-D/Z, file 691

„Außerdem , so Axel Singhofen , Giftstoff-Experte von Greenpeace in Brüssel , sollte die Bestimmung [1 nur sechs Phtalate] betreffen - hauptsächlich **[2 solche]** , [3 die] gar nicht in Spielzeug verwendet werden .“
Engl.: Besides, said Axel Singhofen, expert on toxins for Greenpeace in Brussels, the determination was meant to concern [only six phtalates] - mainly [2 those] [3 which] are not used for toys.

- ⇨ markable 2 = instance; markable 3 = bound
- ⇨ not annotated, because markable 2 cannot be considered as being an instantiation of the class „Phtalate“ which is referred to by the quantified expression.

Example 3.34, TüBa-D/Z, file 815

„Ich könnte der taz auf die Sprünge helfen , denn Böll und ich begründeten einst das Kartell , das weder eins war noch sich in Unterschriften erschöpfte , sondern Solidarität mit Verfolgten leistete , was [1 so mancher Spötter], und **[2 deren]** gab es so viele wie Sandflöhe am Meer , jeweils dann wußte , wenn ihm Ungemach drohte - in solchen Fällen baten die tapfren Schreiberlein beim " Kartell " um Hilfe .“
„I could get the taz going because Böll and I once founded the cartel that neither was one nor was it restricted to signatures; instead it showed solidarity with the haunted which [many a mocker], and [2 of those] there existed as many as jigger fleas by the sea, knew exactly when being in trouble - in these cases the brave little tailors asked the „cartel“ for help.“

- ⇨ markable 2 = instance

Example 3.35, TüBa-D/Z, file 848

„Es gibt [1 2.000 Termitenarten] auf der Welt , **[2 die meisten]** in den Tropen oder Subtropen , selten auch in Trockenzonen .“

Engl.: „There are [1 2.000 species of termites] in the world, [2 most of them] in the tropics and subtropics, rarely also in the arid environment.

⇨ markable 2 = instance

⇨ not annotated because the antecedent is a quantified NP, thus markable 2 is not an instantiation of the entire class.

Example 3.36, TüBa-D/Z, file 852

„[Alle siebzehn AnwältInnen Öcalans] sind Verfechter der kurdischen Sache . **[Viele]** stammen aus dem Umfeld des Menschenrechtsvereins IHD , der durch seine Nähe zur PKK auffiel .“

Engl.: [1 All seventeen advocats of Öcalan] are defenders of the Kurdish case. [2 Many] originate from the periphery of the human rights association IHD which attracted the attention because it is close to the PKK.

⇨ markable 2 = instance

⇨ not annotated because the antecedent is a quantified NP, thus markable 2 is not an instantiation of the entire class.

Example 3.37, TüBa-D/Z, file 854

„" [1 Die Hosen] an " hatte sie nur im übertragenen Sinne , im wirklichen Leben aber trug sie niemals **[2 welche]** .“

Engl.: „“[The trousers] on“[11] she had only in a figurative sense; in real life she never wore [2 any].“

⇨ markable 2 = instance

Example 3.38, TüBa-D/Z, file 929

„Eine Arbeit mit [1 gelegentlichen Sternstunden] . **[2 Eine]** erlebte er , als ihm eine Kundin überschwenglich siebzig Mark Trinkgeld in die Hand drückte .“

Engl.: „A job with [1 occasional highlights]. [2 One] he experienced when a client exuberantly gave him seventy marks tip.“

⇨ markable 2 = instance

Conclusion: the anaphor is not annotated as „instance“ if the antecedent does not concern the whole class of objects but a quantified portion of that class.

3.3.7 Expletive

This category is different from all the other annotation categories described so far. It is an attribute value assigned to the impersonal third person singular pronoun “it” (German “es”) and does not denote a specific type of referential relation. On the contrary, it is used for those cases where the pronoun is non-referential. In the literature, this is also know as “pleonastic IT”. But note that it is not used for event anaphora, i.e. uses of “it” where the pronoun refers to an underlying proposition, as e.g. in “Gasoline prizes are rising again and I do not like [it]”.

[11] Note: German idiom „die Hosen anhaben“ corresponds to engl. „to wear the breeches“

In German, impersonal 3d person sg. pronoun “ES” is used as:

1.) Personal Pronoun:
„[1 Das Baby] liegt in der Wiege. [2 Es] schläft ruhig.“
Engl.: [The baby] is lying in the cradle. [It] is sleeping calmly.“

2.) Subject of weather verbs and verbs with missing agent:
“[1 Es] regnet.“
Engl.: [It] is raining.

„[1 Es] trug ihn aus der Kurve.“
Engl.: [It] threw him out of the bend.

3.) Anticipating pronoun in extraposed sentences:
„[1 Es] ist gut, dass Peter kommen konnte.“
Engl.: [It] is good that Peter could come.

4.) Expletive pronoun in sentence-initial position:
„[1 Es] kamen zwei Männer zur Tür herein.“
Engl.: [It] came two men through the door.

„[1 Es] wurde bis zum Morgen getanzt.“
Engl.: [It] was danced until morning[12].

We decided to use the category for the cases 2 and 4. The category is applied to presentational “es” as in example 3.39

Example 3.39: TüBa-D/Z, file 120

```
‟Die Deutschen sehnen sich zurück: [Es] muss mal wieder Krieg sein.″
Engl.: „The Germans are longing to have sth. back: [It] has to be war again.‟
```

It is used for “es” in sentence-initial position, cf. example 3.40.

Example 3.40:

```
[Es] waren zwei Königskinder.
Engl.: [It] were two royal children.
```

It is also applied to cases where the “es” is a substitute for the missing subject, cf. example 3.41.

Example 3.41: TüBa-D/Z, file 500

```
Dennoch gelingt [es] Christian Ebert mit seiner Inszenierung, dem Briefroman Bühnenleben einzuhauchen.

Neverthless, [it] is managed by Christian Ebert with his production to give stage life to the epistolary novel.
```

[12] The German sentence would correspond to „there was dancing“ in English.

Discussion / Problematic Cases

We compared our approach to that of Boyd et al. (2005) Evans (2000) and Paice/Husk (1987) although idiosyncrasies are common in these cases and some of the phenomena described in these publications are restricted to the English language. On the other hand, many of the German examples cannot easily be transferred to English. Besides, there are also syntactic characteristics included in the annotation of non-referential "it". Consequently, two types of expletive "ES" are already marked in the syntactic annotation of the German treebank: "Vorfeld-ES" and "Korrrelat-ES"[13]. These cases are not annotated separatedly in the process of the annotation of referential relations being inherently defined as "expletive".

In the following, we list some of the cases from the TüBa-D/Z treebank that were discussed and which exhibit the difficulties in translating expletive "ES" to another language:

- "Es findet sich" -> idiomatic, not annotated (engl.: it will work out)
- "Es geht (mir) um..." -> expletive (engl.: it is about...)
- "Es ist mir ernst" -> expletive (engl.: it is serious for me)
- "Es kam dazu..." -> expletive (engl.: it happened that..)
- "Es gibt mehrere Lösungen." -> expletive (engl.: there are several solutions)
- "Es handelt sich um..." -> "Es"=expletive; "sich"=inherently reflexive (not annotated) (engl.: it concerns...)
- "Es geht nach seinem Willen" -> expletive (engl.: it is done according to his will)
- "Es ist Krieg" -> not annotated (engl.: it is war)
- "Es waren 765" -> expletive (engl.: it were 765)
- "Es geht los" -> not annotated (engl.: it is starting)
- "Es faellt schwer -> expletive (engl.: it is hard)
- "Es scheint als ob... -> expletive (engl.: it seems as if...)
- "Es steht schlecht um ihn" -> expletive (engl.: it is bad for him)
- „Es macht nichts" -> not annotated (engl.: it doesn't matter)
- „Es fragt sich..." -> "Es"=expletive; "sich"=inherently reflexive (not annotated) (engl.: it has to be asked..)
- „Es hat sich ergeben, dass.." -> "Es"=expletive; "sich"=inherently reflexive (not annotated) (engl.: it arose that...)
- "Es ist das erste Mal" -> expletive (engl.: it is the first time)
- „Es ist ein Mädchen!" -> cataphoric (engl.: it is a girl)
- „Es ist so, wie es ist." -> first „es" not annotated, second „es" anaphoric (engl.: it is as it is)

[13] for a detailed description see manual on syntactic annotation of the TüBa-D/Z treebank.

4 Conclusion

The precise definition of referential relations and the adaption of the annotation categories to critical cases discussed in the project guarantees a consistent treatment of markables and standardized annotation of the data corpus. Additionally, an Inter-Annotator Agreement was carried out as check for consistency of the annotation. For this purpose, an additional annotator marked-up 70 files out of 766. These files were chosen randomly (file 1, 11,21…) and there existed no advisory exchange between the annotators and other project members. The result of the check can be seen in Versley (2006)

5 References

Boyd, Adriane, Whitney Gegg-Harrison, and Donna Byron (2005): "Identifying non-referential it: A machine learning approach incorporating linguistically motivated patterns." In: *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40-47, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith (2002): "The TIGER treebank." In: Hinrichs, Erhard and Kiril Simov (eds.): Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002) (pp. 24–41). Sozopol, Bulgaria.

Brants, Thorsten (1997). "The NeGra Export Format for Annotated Corpora." Universität des Saarlandes, Computational Linguistics, Saarbrücken, Germany.

Davies, Sarah, Massimo Poesio, Florence Bruneseaux, and Laurent Romary (1998) : « Annotating Coreference in Dialogues: Proposal for a Scheme for MATE." MATE: 1998.

van Deemter, Kees and Rodger Kibble (2000). "On coreferring: Coreference in MUC and related annotation schemes." In: Computational Linguistics, 26(2) (pp. 629–637).

Eckle-Kohler, Judith (1999): „Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora." Berlin: Logos Verlag.

Evans, R. (2001): „Applying Machine Learning Toward an Automatic Classification of *It*." In: the Journal of Literary and Linguistic Computing © Oxford University Press. 16;1. (pp. 45-57).

Fitschen, Arne (2004): „Ein computerlinguistisches Lexikon als komplexes System." In: AIMS Vol. 10 (3) (Doctoral Dissertation, University of Stuttgart).

Hinrichs, Erhard, Sandra Kübler, Karin Naumann, Heike Telljohann, Julia Truschkina (2004): "Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank." In: Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT) (pp. 51-62). Tübingen, December 2004.

Hinrichs, Erhard, Sandra Kübler und Karin Naumann (2005): A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations. In: Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky (pp. 13-20). Ann Arbor, MI, Juni 2005.

Hirschman, Lynette and Nancy Chinchor (1997): "MUC-7 Coreference Task Definition." In: MUC-7 Message Understanding Conference Proceedings. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html (last visited: 24.06.2006, no print version available)

Kunze, Claudia and Lothar Lemnitzer (2002): "GermaNet - representation, visualization, application." In: Proceedings LREC 2002, main conference, Vol V. (pp. 1485-1491).

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993): „Building a Large Annotated Corpus of English: The Penn Treebank.“ In: Computational Linguistics, 19 (2) (pp. 313--330).

Mitkov, Ruslan (ed.) (2003): „The Oxford handbook of computational linguistics“**. -** 1. publ. - Oxford [et.al.] : University Press.

Müller, Christoph and Michael Strube (2003): “Multi-levelannotation in MMAX.” In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan (pp.198-107).

Paice, C.D. and Husk, G.D. (1987): “Towards the Automatic Recognition of Anaphoric Features in English Text: The Impersonal Pronoun 'It'.” In: Computer Speech and Language, 2 (pp.109-132). Academic Press, US

Plaehn, Oliver (1998): „Annotate Bedienungsanleitung“. Universität des Saarlandes, Sonderforschungsbereich 378, Projekt C3, Saarbrücken, Germany, April 2003.

Poesio, Massimo (2000): “Coreference”. In: Mengel, A., Dybkjaer, L., Garrido, J.M., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C. 8. Januar 2000. MATE Dialogue Annotation Guidelines.
http://www.ims.uni-stuttgart.de/projekte/mate/mdag/ (last visited: 24.6.2006)

Poesio, Massimo (2004): “The MATE/GNOME Scheme for Anaphoric Annotation, Revisited.” In: Proceedings of SIGDIAL, Boston, April.
http://cswww.essex.ac.uk/staff/poesio/publications/SIGDIAL04.pdf
(last visited: 24.06.2006)

Stegmann, Rosmary, Heike Telljohann, and Erhard W. Hinrichs (2000): “Stylebook for the German Treebank in VERBMOBIL.” Technical Report 239, Verbmobil.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler and Heike Zinsmeister (2006): „Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).” Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.

Versley, Yannick (2006): „Disagreement Dissected: Vagueness as a Source of Ambiguity in Nominal (Co-)Reference.“ Workshop Paper, ESSLLI 2006.