A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations

Erhard W. Hinrichs, Sandra Kübler, Karin Naumann
SfS-CL, University of Tübingen
Wilhelmstr. 19
72074 Tübingen, Germany
{eh,kuebler,knaumann}@sfs.uni-tuebingen.de

Abstract

This paper reports on the SYN-RA (SYNtax-based Reference Annotation) project, an on-going project of annotating German newspaper texts with referential relations. The project has developed an inventory of anaphoric and coreference relations for German in the context of a unified, XML-based annotation scheme for combining morphological, syntactic, semantic, and anaphoric information. The paper discusses how this unified annotation scheme relates to other formats currently discussed in the literature, in particular the annotation graph model of Bird and Liberman (2001) and the pie-in-the-sky scheme for semantic annotation.

1 Introduction

The purpose of this paper is threefold: (i) it discusses an annotation scheme for referential relations for German that is significantly broader in scope than existing schemes for the same task and language and that also goes beyond the inventory of anaphoric relations included in the pie-in-the-sky sample feature structures[1], (ii) it presents a unified, XML-based annotation scheme for combining morphological, syntactic, semantic, and anaphoric information, and (iii) it discusses how this unified annotation scheme relates to other formats currently discussed in the literature, in particular the annotation graph model of Bird and Liberman (2001) and the pie-in-the-sky scheme for semantic annotation[2].

2 Referential Relations

This section introduces the inventory of referential relations adopted in the SYN-RA project. We define *referential relations* as a cover-term for all contextually dependent reference relations. The inventory of such relations adopted for SYN-RA is inspired by the annotation scheme first developed in the MATE project (Davies et al., 1998). However, it takes a cautious approach in that it only adopts those referential relations from MATE for which the developers of MATE report a sufficiently high level of inter-annotator agreement (Poesio et al., 1999).

SYN-RA currently uses the following subset of relations: *coreferential*, *anaphoric*, *cataphoric*, *bound*, *split antecedent*, *instance*, and *expletive*. The potential markables are definite NPs, personal pronouns, relative, reflexive, and reciprocal pronouns, demonstrative, indefinite and possessive pronouns.

There is a second research effort under way at the European Media Laboratory Heidelberg, which also annotates German text corpora and dialog data with referential relations. Since their corpora are not publicly available, it is difficult to verify their inventory of referential relations. Kouchnir (2003) has used their data and describes the relations *anaphoric*, *coreferential*, *bridging*, and *none*.

Following van Deemter and Kibble (2000), we define a *coreference relation* to hold between two

[1] See e.g. nlp.cs.nyu.edu/meyers/pie-in-the-sky/analysis5.

[2] See nlp.cs.nyu.edu/meyers/pie-in-the-sky/pie-in-the-sky-descript.html.

NPs just in case they refer to the same extra-linguistic referent in the real world. In the following example, a coreference relation exists between the noun phrases [1] and [2], and an *anaphoric relation* between the noun phrase [2] and the personal pronoun [3]. Since noun phrases [1] and [2] are coreferential, all three NPs belong to the same coreference chain. In keeping with the MUC-6 annotation standard[3], we establish the anaphoric relations of a pronoun only to its most recently mentioned antecedent.

(1) [1 Der neue Vorsitzende der Gewerkschaft
The new chairman of the union
Erziehung und Wissenschaft] heißt [2 Ulli
Education and Science is called Ulli
Thöne]. [3 Er] wurde gestern mit 217
Thöne. He was yesterday with 217
von 355 Stimmen gewählt.
out of 355 votes elected.

'The new chairman of the union of educators and scholars is called Ulli Thöne. He was elected yesterday with 217 of 355 votes.'

Cataphoric relations hold between a preceding pronoun and a following antecedent within the same sentence, even if this antecedent has already been mentioned within the preceding text. An example for a cataphoric relation is shown in (2).

(2) Vier Wochen sind [sie] nun schon in Berlin,
Four weeks are they now already in Berlin,
[die 220 Albaner aus dem Kosovo].
the 220 Albanians from the Kosovo.

'They have already been in Berlin for four weeks, the 200 Albanians from Kosovo.'

The relation *bound* holds between anaphoric expressions and quantified noun phrases as their antecedents (see example (3)).

(3) [Niemandem] fällt es schwer, das Bild
To nobody is it difficult, the picture
vor [sich] zu sehen.
in front of himself to see.

'Nobody has trouble imagining the picture.'

[3]See www.cs.nyu.edu/cs/faculty/grishman/COtask21.book_1.html.

The *split antecedent relation* holds between coordinate NPs/plural pronouns and pronouns/definite NPs referring to one member of the plural expression. In example (4), the indefinite pronoun *beide* enters into two split antecedent relations, with noun phrases 1 and 2.

(4) Aber plötzlich gibt es da einen völlig
But suddenly gives it there a completely
unglaubwürdig und grotesk wirkenden
implausible and grotesque seeming
Anruf [1 des Detektiven] bei [2 der
phone call of the detective to the
Mutter des Opfers] , [beide] weinen
mother of the victim , both cry
sich minutenlang etwas
themselves for some minutes something
vor , ...
verb part , ...

'But suddenly, there is a completely implausible and grotesque phone call from the detective to the mother of the victim, they both cry at each other for several minutes, ...'

An *instance relation* exists between a preceding/following pronoun and its NP antecedent when the pronoun refers to a particular instantiation of the class identified by the NP.

(5) Die konservativen Kräfte warten ja nur
The conservative powers wait just only
darauf, ihm [Sätze] um die Ohren zu
for that, him sentences around the ears to
hauen wie [jenen von den 16
hit like the one about the 16
Mittelstrecklern], denen er in vier
middle-distance runners, to whom he in four
Wochen die Viererkette
weeks the double full-back formation
beibringe.
teaches.

'The conservative powers are just waiting to bombard him with sentences like the one about the 16 middle-distance runners who he is teaching the double full-back formation in four weeks.'

In sentence (5), the relation between the two bracketed NPs is an example of such an instance relation since the second NP is a particular instantiation of the referent denoted by the first NP.

A third person singular neuter pronoun *es* is marked as *expletive* if it has no proper antecedent. This is the case for presentational *es* in example (6), impersonal passive as in example (7), or *es* as subject for verbs without an agent as in example (8).

(6) [1 Es] zeichnet sich die konkrete Möglichkeit
It emerges the concrete possibility
ab.
verb part.
'The concrete possibility emerges.'

(7) [Es] wird bis zum Morgen getanzt.
There is until the morning danced.
'People are dancing until morning.'

(8) [Es] steht schlecht um ihn.
It stands bad for him.
'He is in a bad way.'

Apart from expletive uses of *es* and anaphoric uses with an NP antecedent, the pronoun *es* can also be used in cases of event anaphora as in sentence (9). Here *es* refers to the event of Jochen's winning the lottery. Currently, the annotation in SYN-RA is restricted to NP anaphora and therefore event anaphors such as in sentence (9) remain unannotated for anaphora.

(9) Jochen hat im Lotto gewonnen. Aber er
Jochen has in the lottery won. But he
weiss es noch nicht.
knows it yet not.
'Jochen has won the lottery. But he does not know it yet.'

The annotation of such relations is performed manually with the annotation tool MMAX (Müller and Strube, 2003). Its graphical user interface allows for easy selection of the relevant markables and the accompanying relation between the contextually dependent expression and its antecedent.

3 Automatic Extraction of Markables and of Semantic Information

Annotation of referential relations involves two main tasks: the identification of markables, i.e., identifying the class of expressions that can enter into referential relations, and 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. If referential relations need to be annotated from plain text, then markables must be identified semi-automatically from the output of a chunker or full parser, if available, or otherwise 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 required depends on the quality of the parser, but will require at least some amount of manual post-correction 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 (Brants et al., 2002) treebank and the Tübingen treebanks for spoken and written German (Stegmann et al., 2000; Telljohann et al., 2003). All the treebanks were annotated with the help of the annotation tool Annotate (Plaehn, 1998). The treebank annotations are available in the Annotate export format (Brants, 1997) and in an XML format.

The SYN-RA project 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* (*die tageszeitung*). The treebank currently comprises appr. 15 000 sentences, with a new release of 7 000 additional sentences scheduled for June of this year.

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 properties for each markable. The TüBa-D/Z annotation scheme distinguishes four levels of syntactic constituency: the lexical level, the phrasal level, the level of topological fields, and the clausal level. The primary ordering principle of a clause is the inventory of topological fields, which characterize the word or-

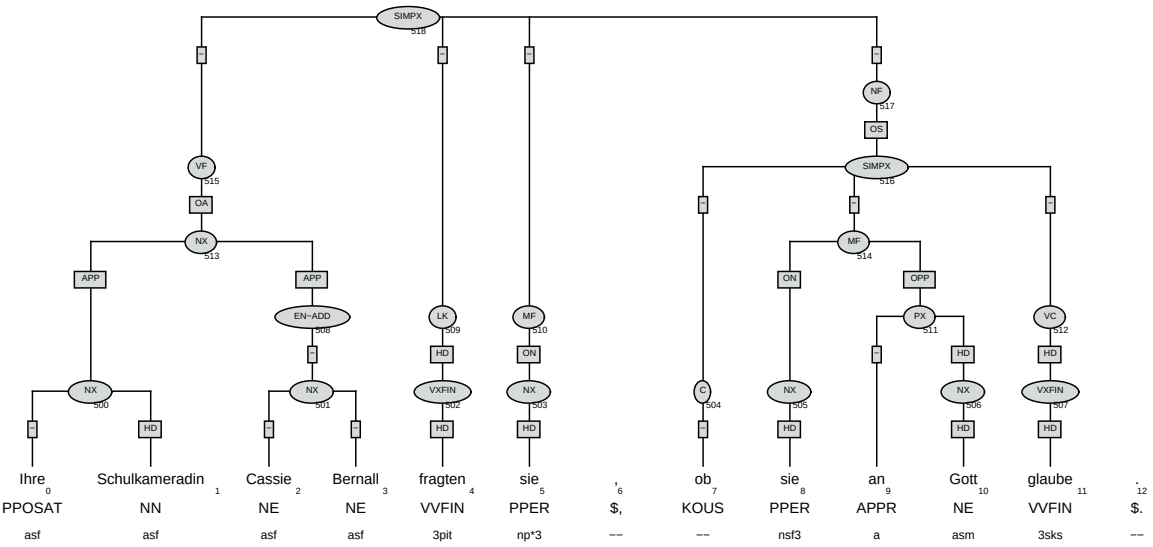


Figure 1: A sample tree from the TüBa/D-Z treebank.

der regularities among different clause types of German and which are widely accepted among descriptive linguists of German (cf. e.g. (Drach, 1937; Höhle, 1986)). The TüBa-D/Z annotation relies on a context-free backbone (i.e. proper trees without crossing branches) of phrase structure combined with edge labels that specify the grammatical function of the phrase in question.

Figure 1 shows an example tree from the TüBa-D/Z treebank for sentence (10). The sentence is divided into two clauses (SIMPX), and each clause is subdivided into topological fields. The main clause is made up of the following fields: VF (mnemonic for: *Vorfeld* – 'initial field') contains the sentence-initial, topicalized constituent. LK (for: *linke Satzklammer* – 'left sentence bracket') is occupied by the finite verb. MF (for: *Mittelfeld* – 'middle field') contains adjuncts and complements of the main verb. NF (for: *Nachfeld* – 'final field') contains extraposed material – in this case an indirect yes/no question. The subordinate clause is again divided into three topological fields: C (for: *Komplementierer* – 'complementizer'), MF, and VC (for: *Verbalkomplex* – verbal complex). Edge labels are rendered in boxes and indicate grammatical functions. The sentence-initial NX (for: *noun phrase*) is marked as OA (for: *accusative complement*), the pronouns *sie* in the main and subordinate clause as ON (for: *nominative complement*).

(10) Ihre Schulkameradin Cassie Bernall fragten
Their fellow student Cassie Bernall asked
sie , ob sie an Gott
they[subj] , whether she[subj] in God
glaube.
believes.
'They asked their fellow student Cassie Bernall whether she believes in God.'

Topological field information and grammatical function information is crucial for anaphora resolution since binding-theory constraints crucially rely on sentence-structure (if the binding theory principles are stated configurationally (Chomsky, 1981)) or on argument-obliqueness (if the binding theory principles are stated in terms of argument structure, as in (Pollard and Sag, 1994)). In the case at hand, the subject pronoun of the main clause, *sie*, cannot be anaphorically related to the object NP *Ihre Schulkameradin Cassie Bernall* since they are co-arguments of the same verb. However, the possessive pronoun *ihre* and the subject pronoun *sie* of the subordinate clause, can be and, in fact, are anaphorically related, since they are not co-arguments of the same verb. This can be directly inferred from the treebank annotation, specifically from the sentence structure and the grammatical function information

encoded on the edge labels. Most published computational algorithms of anaphora resolution, including (Hobbs, 1978; Lappin and Leass, 1994; Ingria and Stallard, 1989), rely on such binding-constraint filters to minimize the set of potential antecedents for pronouns and reflexives.

As already pointed out, the sample sentence contains four markables: one possessive pronoun *Ihre*, two occurrences of the pronoun *sie* and one complex NP *Ihre Schulkameradin Cassie Bernall*. The latter NP is a good example of SYN-RA's longest-match principle for identifying markables. In case of complex NPs, the entire NP counts as a markable, but so do its subconstituents – in the case at hand, particularly the possessive pronoun *ihre*. All of this information can be directly derived from the treebank account. Compared to other annotation efforts for German where markables have to be chosen manually (Müller and Strube, 2003), manual annotation in the SYN-RA project can, thus, be restricted to the selection of the appropriate referential relations between referentially dependent expressions and their nominal antecedents.

4 The Unified, XML-based Annotation Scheme

The annotation of referential expressions is embedded in a unified format which also contains morphological, syntactic, and semantic information. The annotation scheme is represented in XML, the widely acknowledged standard for exchanging data, which guarantees portability and re-usability of the data. Each sentence, as well as all words and all nodes in the syntactic structure, are assigned a unique ID. These IDs are used in the annotation of referential relations. The annotation of the treebank sentence 11976 (cf. example (10)) is shown in Figure 2.

The sentence number is encoded as the ID of the sentence. The first word, *Ihre*, has an anaphoric relation to a noun phrase in the previous sentence. This relation is marked in the element *anaphora*, which gives the antecedent as node 517 of sentence 11975, i.e. the previous sentence. The other two anaphoric relations are sentence-internal, the first personal pronoun *sie* having *Ihre* (id: s11976w0) as antecedent, the second one the noun phrase *Ihre Schulfreundin Cassie Bernall* (id: s11976n513). The annotation of the first personal pronoun is an example for the annotation of an anaphoric chain. *Ihre* and *sie* belong to the same chain. However, in order to facilitate the extraction of direct relations, such chains are represented in a way that each anaphoric expression refers to the last occurrence of an antecedent.

The SYN-RA scheme is very similar to the MUC-6 coreference annotation scheme[4] but it is more powerful in two respects: As described above, the inventory is not restricted to coreference and anaphoric relations, it also covers e.g. instance relations or split antecedent relations. The latter relation is also the reason for encoding the relational information as XML elements, and not as attributes of a word or a node. If an anaphor enters into a split antecedent relation, it has more than one distinct antecedent. In this case, the element *anaphora* has two (or more) relations. Such an example is graphically displayed for sentence (4) in Figure 3. The relevant XML representation of the complex entry for the word *beide* is shown in Figure 4.

5 Related Work

This section discusses how the unified SYN-RA annotation scheme relates to other formats currently discussed in the literature, in particular the pie-in-the-sky scheme for semantic annotation[5] and the annotation graph model of (Bird and Liberman, 2001). While these two annotation schemes are by no means the only contenders for corpus annotation standards in the literature, they are certainly among the most ambitious and promising.

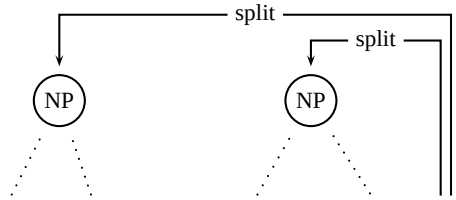
While the pie-in-the-sky scheme is clearly still under development, the following characteristics and goals can already be gleaned from its webpage and the annotation examples presented there: The annotation is feature-structure-based and incorporates various levels of linguistic annotation, in particular a PROPBANK style predicate-argument structure, dependency style syntactic information, as well as morpho-syntactic and word class information. All this information is rooted in the attributes needed for predicate-argument assignment,

[4] See www.cs.nyu.edu/cs/faculty/grishman/COtask21.book_1.html.

[5] See nlp.cs.nyu.edu/meyers/pie-in-the-sky/pie-in-the-sky-descript.html.

```
<sentence  id="s11976">
  <node id="s11976n518" cat="SIMPX" func="--" parent="0">
    <node id="s11976n515" cat="VF" func="-">
      <node id="s11976n513" cat="NX" func="OA">
        <node id="s11976n500" cat="NX" func="APP">
          <word id="s11976w0" form="Ihre" pos="PPOSAT" morph="asf" func="-">
            < anaphora>
              < relation type="ana" antecedent="s11975n517"/>
            < /anaphora> </word>
          <word id="s11976w1" form="Schulkameradin" pos="NN" morph="asf" func="HD"/>
        </node>
        <node id="s11976n508" cat="EN-ADD" func="APP">
          <node id="s11976n501" cat="NX" func="-">
            <word id="s11976w2" form="Cassie" pos="NE" morph="asf" func="-"/>
            <word id="s11976w3" form="Bernall" pos="NE" morph="asf" func="-"/>
          </node> </node> </node> </node>
    <node id="s11976n509" cat="LK" func="-">
      <node id="s11976n502" cat="VXFIN" func="HD">
        <word id="s11976w4" form="fragten" pos="VVFIN" morph="3pit" func="HD"/>
      </node> </node>
    <node id="s11976n510" cat="MF" func="-">
      <node id="s11976n503" cat="NX" func="ON">
        <word id="s11976w5" form="sie" pos="PPER" morph="np*3" func="HD">
           < anaphora>
            < relation type="ana" antecedent="s11976w1"/>
          < /anaphora> </word> </node> </node>
    <word id="s11976w6" form="," pos="$," morph="--" func="--" parent="0"/>
    <node id="s11976n517" cat="NF" func="-">
      <node id="s11976n516" cat="SIMPX" func="OS">
        <node id="s11976n504" cat="C" func="-">
          <word id="s11976w7" form="ob" pos="KOUS" morph="--" func="-"/>
        </node>
        <node id="s11976n514" cat="MF" func="-">
          <node id="s11976n505" cat="NX" func="ON">
            <word id="s11976w8" form="sie" pos="PPER" morph="nsf3" func="HD">
               < anaphora>
                 < relation type="ana" antecedent="s11976n513"/>
               < /anophora> </word> </node>
          <node id="s11976n511" cat="PX" func="OPP" comment="">
              <word id="s11976w9" form="an" pos="APPR" morph="a" func="-"/>
              <node id="s11976n506" cat="NX" func="HD">
                <word id="s11976w10" form="Gott" pos="NE" morph="asm" func="HD"/>
              </node> </node> </node>
        <node id="s11976n512" cat="VC" func="-">
          <node id="s11976n507" cat="VXFIN" func="HD">
            <word id="s11976w11" form="glaube" pos="VVFIN" morph="3sks" func="HD"/>
          </node> </node> </node> </node> </node>
   <word form="." pos="$." morph="--" func="--" parent="0"/>
</sentence>
```

Figure 2: The XML format represents information on all levels of annotation. The words of the sentence and the anaphoric annotation are shown in bold.



Aber plötzlich gibt es da einen ... Anruf **des Detektiven** bei **der Mutter ..., beide** weinen sich minutenlang etwas vor ...

Figure 3: The annotation of the split antecedent relation in sentence (4). For representational reasons, the sentence is shortened and only relevant information is displayed. Syntactic boundaries are shown as dotted lines, anaphoric relations as black lines.

```
<word id="s3426w20" form="beide" pos="PIS" morph="np*" func="HD">
   < anaphora>
      <relation type="split" antecedent="s3426n507"/>
      <relation type="split" antecedent="s3426n526"/>
   < /anaphora>
</word>
```

Figure 4: The XML representation of the encoding of split antecedents for the word *beide* in sentence (4). A graphical representation of the relation is shown in Figure 3. The antecedent "s3426n507" refers to the first NP, "s3426n526" to the second one in Figure 3.

with syntactic and morpho-syntactic information distributed among the corresponding elements in the predicate-argument structure representation. Accordingly, semantic representations provide the organizing principle while morpho-syntactic and syntactic information play a subordinated role.

The SYN-RA annotation scheme resembles the pie-in-the-sky scheme in that it also uses one level of representation, in this case hierarchical syntactic structure, as the organizing principle and treats referential relations, grammatical function information, and morpho-syntactic annotation as subordinated types of information. More generally, the pie-in-the-sky and the SYN-RA representations offer a particular view of the annotation, each with its own "perspective": semantics-based (pie-in-the-sky) and syntax-based (SYN-RA).

By contrast, Bird and Liberman's (2001) annotation graphs are intended as a graph-based, multi-layered annotation scheme where each level of linguistic annotation is treated equally, as an independent layer. The graph-based annotation model is powerful enough to also allow groupings of discontinuous constituents and other non-adjacent linguistic phenomena, without having to rearrange the linear order of the input. In both respects, their annotation model is maximally general.

6 Future Directions

In the previous section we have compared two perspective-dependent annotation schemes that use a particular level of linguistic annotation as their primary organizing principle and have contrasted them with the perspective-independent annotation-graph model. We believe that both types of representation models have their independent justification. Perspective-based representations, such as SYN-RA and pie-in-the-sky, are well-justified for particular application scenarios. For example, for text summarization and other semantic tasks, the pie-in-the-sky model seems particularly well-motivated since the pertinent semantic information can be easily extracted from its predicate-argument-structure-rooted feature structures. For other tasks, such as anaphora resolution, for which syntactic information is more relevant, the syntax-based representation of SYN-RA allows for an easier extraction of the relevant information for rule-based, statistical,

and machine-learning approaches to computational anaphora resolution. More generally, perspective-based representations are highly task-dependent. It would be misguided to consider them as ideal, task-independent annotation standards. If one wants to establish a task-independent annotation standard, then a perspective-independent annotation scheme such as the annotation graph model looks like a promising direction for future research. In particular, such research should focus on techniques that allow for easy conversion of perspective-independent representations to task-dependent views of the relevant linguistic information.

References

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60.

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

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

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

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

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

Erich Drach. 1937. *Grundgedanken der Deutschen Satzlehre*. Diesterweg, Frankfurt/M.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Tilman Höhle. 1986. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

Robert J. P. Ingria and David Stallard. 1989. A computational mechanism for pronominal reference. In *Proceedings of the 27th Conference of the Association for Computational Linguistics*, pages 262–271, Vancouver, Canada.

Beata Kouchnir. 2003. A machine learning approach to German pronoun resolution. Master's thesis, School of Informatics, University of Edinburgh.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Christoph Müller and Michael Strube. 2003. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan.

Oliver Plaehn, 1998. *Annotate Bedienungsanleitung*. Universität des Saarlandes, Sonderforschungsbereich 378, Projekt C3, Saarbrücken, Germany, April.

Massimo Poesio, Florence Bruneseaux, and Laurent Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Proceedings of the ACL Workshop on Standards for Discourse Tagging*, pages 65–74.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. University of Chicago Press, Chicago, IL.

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

Heike Telljohann, Erhard W. Hinrichs, and Sandra Kübler, 2003. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.