A Pilot Study on Computer-aided Coreference Annotation

Yannick Versley, Holger Wunsch, Heike Zinsmeister
SFB 441 – Project A1
University of Tübingen
Wilhelmstr. 19
D-72074 Tübingen
Germany
{versley,wunsch,zinsmeis}@sfs.uni-tuebingen.de

Abstract

We present the results of a pilot study on increasing the efficiency of coreference annotation by integrating the predictions of existing coreference components. While similar approaches are already quite common for other linguistic annotation tasks, our experiments are the first to address a more complex task such as coreference annotation.

1 Introduction

An overwhelming majority of currently successful approaches to problems in the domain of Natural Language Processing (NLP) are data-driven, either in using machine learning to automatically learn the parameters of a model, or by using corpora to validate or tune hand-written algorithms. A necessary precondition for such approaches is the presence of an annotated corpus of adequate size, and the cost of their creation does have a very noticeable impact on the total cost of creating NLP components for a given language and/or domain. This cost, due to the manpower needed both for the annotation and for the development of annotation guidelines and, possibly, annotation tools, can usually be justified by the longevity of such resources.

Since the creation of large referentially annotated corpora guided by linguistic principles is a comparatively younger development, we will illustrate some of the driving forces by taking a glance at treebanks, as they have been around significantly longer than other large annotated corpora. As an example, the Penn treebank (Marcus et al., 1993) has been in use for the data-driven building of parsers for quite a long time, from Magerman (1995) to recent approaches like the one of Charniak and Johnson (2005), who use different features and get qualitatively better results. There is even research that uses the Penn Treebank data to induce linguistically richer structures than directly available in the Penn Treebank (Hockenmaier and Steedman, 2002; Cahill et al., 2002; Miyao and Tsujii, 2005).

To improve both the speed and the quality of the annotation process, it is common to develop specialised annotation tools that allow for greater ease of use and for easier consistency checking. In addition to this, it is possible to use the predictions of an automated system to ease the otherwise tedious and time-consuming task of annotation.

In this paper we present the results of a pilot study that examines the effects of automatic pre-annotation on the performance of human annotators. In particular, we report on experiments conducted on marking referential relations in a treebank of German newspaper text. To this end, we use the output of an automatic system for resolving nominal anaphora to pre-annotate data which is afterwards edited manually using specialized annotation software. Our expectations are that guided annotation can both reduce the time needed for annotation, as well as decrease the rate of annotation errors.

1.1 Related work

Given text that is partially (pre-)annotated with the structures that an automatic system predicts, the task of the annotator is reduced to checking and modifying existing annotations as opposed to creating everything from scratch. This approach is especially suited for tagging tasks like part-of-speech tagging or named-entity annotation (e.g., for bootstrapping an annotated corpus in a new domain), where the annotation is simple and easily manageable. For annotation tasks that create more structure, this approach has been followed for syntax in the Penn Treebank (Marcus et al., 1993), where partial parses from a rule-based parser were used (leading to significantly greater productivity than in the creation of Susanne, a similar treebank that has been annotated completely by hand, cf. Sampson, 1993).

Pre-annotation is not the only possible scenario to improve the productivity of the annotation process, as tighter integration is possible: either in guiding or selecting structures built by an automatical system (Kaplan and Maxwell, 1996; van der Beek et al., 2002; Zinsmeister et al., 2002), which usually also means that the annotation process is radically different from the normal annotation process, or by providing editing operations that use an operational model to simplify interaction in clear-cut cases (“Do what I mean” or DWIM operations). In the domain of syntactic annotation, the *annotate* tool (Brants and Plaehn, 2000) uses a statistical model of phrase structure to set the constituent label for newly created phrases, and the

XCDG tool for dependency annotation (Foth et al., 2004) is able to make use of a constraint dependency grammar to change the dependency label, the dependency head or the lexical entry to the value ranked most highly by the grammar, speeding up annotation in numerous cases.

The most commonly voiced concern against pre-annotation is the bias towards system decisions that it entails, as annotators would leave any plausible suggestion intact, hiding ambiguities that would otherwise become apparent. This is especially important for both calculating annotator agreement, as ambiguities would then not lead to disagreement, and system evaluation, as annotators will more often err in favor of the system than against it.

However, the potential benefits of simplifying interaction by pre-annotation or deeper integration approaches are twofold: firstly, the overall annotation speed increases since fewer and/or simpler interaction is required. Secondly, an overall increase in annotation quality could be hoped for: the kinds of errors produced by an automatic preprocessor are different from the kinds of errors procduced by humans. Given that the annotator applies the guidelines correctly, the remaining errors introduced are of random nature and frequency. With pre-annotated data, the majority of the decisions to be made is reduced to simply checking the automatic suggestions. If the output of the preprocessor is of sufficient quality, this can reduce the error rate. Furthermore it is expected that the kinds of errors in the pre-annotated data are predictable with proper knowledge of the system's bias.

To this date, approaches of automatic pre-annotation have been mainly used in POS-tagging and syntax annotation. To our knowledge, the present study is the first to apply automatic pre-annotation to the task of coreference annotation.

2 Coreference Annotation

In this section we describe the overall setting of coreference annotation of the Tübingen Treebank of Written German (TüBa-D/Z) (Hinrichs et al., 2004) which serves as the data source of our experiments.

In general, the basis of linguistic annotation is a well-defined annotation scheme, which normally is a compromise between (1) descriptive adequacy, ideally following some comprehensive theory that explains the phenomena; (2) specifiability of the phenomena: whether human annotators can recognise the phenomena in text and disambiguate between alternative interpretations; (3) suitability for the intended application: the granularity of the annotation scheme is influenced by a utility-cost trade-off that takes into account the requirements of the target application and also its limitations. The process of annotation might become too time-consuming and costly if an annotation scheme is more fine-grained than needed for the target application.

An additional criterion to be taken into account is standardisation. To ensure comparability and reusability, it is essential to conform to some standard (cf. Zinsmeister et al. 2007). For this reason, the set of referential relations that we use is strongly inspired by the annotation scheme first developed in the MATE project (Poesio 2000; for a refinement see Poesio 2004) which serves as a standard and is itself based on an evaluation of five predecessors: MUCCS (Hirschman, 1997), the DRAMA scheme (Passoneau, 1997), the UCREL scheme (Fligelstone, 1992), the scheme developed by Bruneseaux and Romary (1997) and the MapTask landmarks (Anderson et al., 1991). We make use of a subset of the relations that are propsed by MATE.

2.1 Set of Relations

The major goal of our project is to create tools for pronoun resolution and for resolution of definite noun phrases. For reasons of feasibility the domain is restricted to nominal elements in the text, i.e., only pronominals and noun phrases are annotated both as anaphors as well as antecedents. We mark neither event anaphora nor zero anaphora.

The core relations concern *identity-of-reference anaphora* which means that the anaphor and its antecedent refer to the same extra-linguistic referent.

Our annotation scheme is based on the MATE recommendations (Poesio, 2000). In some cases we deviate from the original MATE label which is then given in parenthesis. In what follows, we illustrate the relations which are preprocessed in our experiment: *coreferential*, *anaphoric*, and *cataphoric.*

2.1.1 Coreferential

Two noun phrases that refer to the same referent in some mental space. Number, gender and case mismatches are irrelevant. Markables 1 and 2 are coreferent:
Sweden and [1 Finland] in the quarter final of the ice-hockey world cup: 6:1 score against Switzerland was the second victory of the title-holder, which has 4:0 points just like [2 the Fins] (4:1 over Belarus).[1]

2.1.2 Anaphoric

A set relation between a definite pronoun (including reflexives) and its antecedent in the preceding text. Markables 1, 2, 3, 4, and 5 all belong to the same reference set:
[1 A clear sound] spreads [2 itself] out, warm and full, until [3 it] fills the whole hall. After that, [4 it] thins out, falls into pieces and fades [5 itself].[2]

Lexicalized reflexives that belong to phrasal verbs do not enter an anaphoric relation. There are heuristics for the annotators to identify these exceptions:
Er beeilt sich. (*he hurries*)

[1] Translated from TüBa-D/Z, sentence 4757: *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 hatte wie die Finnen (4:1 über Weißrussland).*

[2] Translated from TüBa-D/Z, sentence 6831f.: *[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].*

2.1.3 Cataphoric

A set relation between a pronoun and a local antecedent that follows in the same sentence.[3] Markable 1 is cataphoric to markable 2:
Als "Open Air Festivalorchester" lädt das in [1 seiner] Existenz gefährdete [2 Rundfunkorchester Berlin] am 18. Juli zu einem Strauß-Konzert vor dem Französischen Dom ein.

(*The [2 Broadcasting Orchestra Berlin], which is endangered in [1 its] existence, invites to a Strauß-concert in front of the French Dome on the 18th of July under the name of "Open Air Festival Orchestra"*)

In addition to the three core relations, our annotation scheme provides other coreference relations that do not conform to the identity-of-reference description: *split_antecedent*, *instance* and *bound*, and, additionally, a label for expletive pronouns. As they are ignored in the experiments, we will not elaborate on them any further (see Naumann, 2006).

3 The PALinkA annotation software

The software that we use for annotating referential relations in our project is the PALinkA annotation tool (Orasan, 2005), which is developed at the University of Wolverhampton (UK). We also deployed PALinkA in the experiments we report on in this paper. PALinkA allows an annotator to define spans of text as markables[4] and then add referential relations between them. The tool uses different colors to highlight markables and relations, so it is fairly easy for the annotator to get an overview of where markables are and what referential relations hold between them.

The data model that PALinkA assumes for referential relations is that of *coreference chains*. All coreferent markables become part of a chain within which any markable is linked only to its direct antecedent. Other annotation tools, such as MMAX (Müller and Strube, 2003) use a *coreference set* model of data representation. In this view, coreferent markables are members of a set instead of being interlinked. The technical advantage of the latter approach is that for any two markables, it is immediately possible to find out whether they are coreferent. The advantage of the chain model is that it is a more natural representation of how annotators envision coreference when annotating - primarily as a relation between two markables.

The set of coreference relations to be used in an annotation project is fully configurable. PALinkA has a built-in timer that measures per-article annotation time. The tool uses an XML-based format for storing coreference data. It is an inline format, which means that words, markables and coreference relations are all kept in the same file.

4 Unguided versus guided annotation

The annotation of referential relations is usually a two-step process of identifying markables and then annotating the referential relations proper.

Markable identification can be completely unrestricted (*unguided*), which allows the human annotator to freely define sequences of words that he or she regards as a markable. The advantage of this is that the annotator is able to decide what the markable should actually be on a case-by-case basis. However, this comes at a price. The fact that the annotator first has to find the relevant markables means that the annotation process takes more time. More important however is that the manual process of identifying markables introduces a non-neglectible source of error. This makes it hard to align the annotation with other layers of annotation of the same data that frequently exist, or with automatically pre-processed data which is to be merged with the coreference information.

The approach taken in the TüBa-D/Z annotation project is a semi-automatic, or *guided*, strategy of annotation, which remedies the problems discussed above. Instead of leaving it up to the annotator, markables are automatically extracted from the corpus. The extraction is based on a strict syntactic definition: All phrases with category NX (noun phrases), all attributive possessive pronouns with part of speech PPOSAT, and all attributive relative pronouns with part of speech PRELAT are markables. No additional markables are added manually. The advantage of this approach is that the annotation is guided such that the annotator remains consistent with the linguistic design descisions, and the coreference annotation is structurally compatible with the other layers of annotation in the treebank and can readily be merged with the corpus.

5 Data and experimental setup

For the experiments, we combined two different systems for the automatic resolution of referential relations. The first system concentrates on finding referential relations between definite noun phrases (relation *coreferential*), while the second system resolves third person pronouns to their NP or pronominal antecedents (relations *anaphoric* and *cataphoric*). For the remaining types of relations, no automatic suggestions are generated.

5.1 The data

The basis for the experiments is the TüBa-D/Z corpus (Telljohann et al., 2006) that consists of about 33,000 sentences of German daily newspaper text. The TüBa-D/Z has been annotated manually and contains a layer of part-of-speech and morphological information, a chunk layer, a layer of topological fields (a widely accepted analysis of German sentence structure), a clausal layer, and finally a layer of coreferential relations.

[3] Nominals in *as*-phrases do not qualify as markables.

[4] We do not use this feature in our project because markables are automatically extracted from the syntactic annotation layer of TüBa-D/Z

	Prec	Recl	F
Resolution of definite NPs			
unmodified system	61.6	70.1	65.6
only *coref-yes*	80.2	54.5	64.9
all suggestions	59.3	72.4	65.2
Resolution of pronouns			
full evaluation data	76.3	96.6	85.2
first 125 articles	78.8	96.7	86.9

Table 1: *Results for definite descriptions, names, and pronouns*

5.2 Resolution of definite noun phrases

The module for the coreference resolution of names and definite noun phrases is based on the approach of Versley (2006)[5]: a first module looks for a morphologically compatible candidate that shares the same lexical head and has compatible modifiers, in a fashion similar to the one described by Vieira and Poesio (2000). If both mentions contain names, it is required that the names match; this eliminates false positives where an appositive noun matches, such as *Mr. Putin, president of the Russian Federation* and *President Clinton.* For definite descriptions that have no same-head antecedent and are not identified as part of an idiomatic construction, several sources of semantic and syntactic information are used to find plausible antecedents or rule out resolution for discourse-new definites and associative bridging anaphora:

- A coarse semantic classification, based on the GermaNet hierarchy (Hamp and Feldweg, 1997) and named-entity classification.
- The grammatical function and number of the definite description, which is often indicative for the discourse-new/discourse-old distinction
- Node distance in GermaNet
- A predicate compatibility statistic (Versley, 2006), which allows to rule out anaphor-antecedent pairs by coarse modeling of verbs' selectional preferences. This is also helpful in the case of ambiguous or less frequent nouns, as verbs contribute some information.
- An approach for combining frequency counts of several patterns indicating an instance relation on the World Wide Web, as described in Versley (2007). Using pattern search is a useful source especially for anaphoric reference to named entities, which are too numerous to be covered well in GermaNet.

In difference to the use in fully automatic coreference annotation, the variant we used here tries to make a compromise between high precision (essential for keeping annotators from mis-identifying antecedents) and high recall (essential for letting annotators detect discourse-old noun phrases). To reach this compromise, we have integrated a measure of certainty into the system, and according to this, we label the relations using either the *coref-yes* label (for good certainty) or the *coref-maybe* label (for weak certainty), allowing the annotator to conveniently look for a plausible antecedent in the positive case but also making clear uncertain relations, which could well be false positives.

Feature	Values
Type of pronoun	reflexive, possessive, personal
Type of relation	anaphoric, cataphoric
Equality of grammatical function (GF)	same, different
Sentence distance	same clause, $0 \ldots 3$
Word distance	$0 \ldots n$
GF of pronoun and antecedent	(all GFs in TüBa-D/Z)
Type of NP	proper, common
NP definiteness	definite, indefinite, n/a
NP nesting	top, embedded

Table 2: *Features used in the memory-based pronoun resolver*

5.3 Resolution of third person pronouns

The system for pronoun resolution (Hinrichs et al., 2005) deals with third person reflexive, possessive, and personal pronouns. It adopts a hybrid architecture of three stages. The first stage is a morphological prefilter that rules out pairs of a pronoun and a candidate antecedent that do not agree in number and gender. The core module in the second stage is built upon the TiMBL memory-based classifier (Daelemans et al., 2005). Pronoun resolution is reformulated as a binary classification task with the aim of deciding whether there exists an anaphora or cataphora relation between a pronoun and a candidate antecedent. The features that are presented to the classifier are listed in table 2. Finally, as the third stage, a set of rule-based postfilters is applied to the output. For pronouns that the classifier found multiple antecedents for, only the left-most one is picked[6]. If a pronoun could not be resolved at all, the closest morphologically compatible subject is selected as the antecedent. The system is evaluated using ten-fold cross-validation, results are listed in table 1.

5.4 Data set

For the experiments, we chose a subset of 20 articles in PALinkA-XML format, which had already been annotated by two annotators in the course of normal annotation (see table 3), of roughly balanced size between 10kB and 15kB from the TüBa-D/Z. The number of words per article ranges between 96 and 495, with an average of 291 words. There is a total of 2,329 markables in the article files. Of each of these 20 articles, we generated two files in the XML-based native PALinkA format: One file contains markables only, but remains

[5] Evaluation results are given in table 1.

[6] for German, this works better than picking the closest antecedent

1	T920711.82	11	T920711.160
2	T920711.105	12	T920711.163
3	T920711.115	13	T920711.187
4	T920711.116	14	T920711.190
5	T920711.117	15	T920711.203
6	T920711.129	16	T920711.210
7	T920711.134	17	T920711.211
8	T920711.139	18	T920711.212
9	T920711.148	19	T920711.213
10	T920711.152	20	T920711.214

Table 3: *The articles used in the experiments*

unannotated otherwise. The second file was created by separately running the NP and pronoun annotation systems, and afterwards merging the output of both systems into one file. The 40 files produced this way constitute the experiment data set.

The NP coreference resolver found 197 coreference relations, of which it judged 126 as certainly correct, and 71 as maybe correct.

The pronoun resolver annotated 96 relations, of which 78 are anaphoric and 17 cataphoric. The pronoun resolver does not assign a confidence measure to its annotations.

5.5 The Annotators

A total of four annotators was involved in the experiments. All of them were student assistants working in our research project, and native speakers of German. Two of the student assistants (annotators A and B) were fairly new to annotating referential relations and had joined the project and undergone a period of hands-on annotation training just recently before the experiments started.

Annotator C already had about one year of experience of annotating in the project. Annotator D started at the same time as annotator C, but had left the project at the time of the experiments.

The experiment data set had already been annotated by annotators C and D several months earlier (but at a time when they had been on the project long enough to be considered experienced annotators). We use this data set as a reference annotation, and refer to it by C_{ref} and D_{ref}.

As part of the experiments, annotator C was asked to work on the same data again. With several months having passed between the annotation of the reference data and the annotation of the data for the experiments, we can exclude falsifying effects on the results that might otherwise have arisen from annotator C remembering parts of the data.

5.6 Experiments

The objective of the experiments was to find out whether the automatically generated suggestions are influential on the human annotation process with respect to both quality and annotation time.

The annotators used the PALinkA software for both the plain article files and the article files that contain suggestions. When annotating plain articles, the annotators conducted the annotation process as usual.

Automatic	**Manual**
suggest_coref_yes	*coreferential*
suggest_coref_maybe	
suggest_anaphoric	*anaphoric*
suggest_cataphoric	*cataphoric*

Table 4: *Categories of automatic and manual referential annotation*

As of the current version of PALinkA, there is no dedicated feature for computer-aided coreference annotation. Therefore we just added four new relation types for *suggest_coref_yes*, *suggest_coref_maybe*, *suggest_anaphoric*, and *suggest_cataphoric* (see table 4), and chose appropriate colors for highlighting the new relations in the PALinkA relations editor. This allowed the annotators to easily identify the suggested relations as such. Whenever they found an incorrect suggestion, they could either override it with the correct relation (using the "non-suggestion" variant of the relation), or delete the incorrect relation altogether. Suggestions that were resolved to the right antecedent were left unchanged, and the relation labels were replaced afterwards with their unmarked counterparts (that is *suggest_coref_yes* and *suggest_coref_maybe* become *coreferential*, *suggest_anaphoric* becomes *anaphoric*, and *suggest_cataphoric* is mapped to *cataphoric*; see table 4).

We used the timing function that is built into PALinkA to measure how long it took an annotator to annotate an article. The timing values accumulate when articles are multiply edited.

We conducted the experiment in two stages. In stage 1, all three annotators were asked to annotate the same set of four files, as shown in table 5. The first two files were unannotated, the second pair of files contained suggestions. The purpose of this first stage was for us to be able to measure and compare the individual differences in annotation quality and speed.

In stage 2, the annotators worked on the remaining 16 articles. Annotators A and B were presented with alternately one plain article and one pre-annotated article, such that whenever annotator A worked on a plain article, annotator B got a pre-annotated article, and vice versa (see table 5). By making the annotators annotate an article only once, we made sure that the timing data did not get distorted because the annotators remembered the content of the article. Annotator C had already annotated the 16 remaining plain articles several months earlier. Therefore, the file set annotator C was presented with only contained pre-annotated articles.

5.7 Agreement and Speed Results

We computed two agreement statistics, the partition-based F-measure of Vilain et al. (1995) and the partition-based kappa statistic of Passonneau (1997). Both measures were calculated between annotators and a reference annotation of the articles, which was

Annotator:	**A**	**B**	**C**
Stage 1			
T920711.82	plain	plain	plain
T920711.105	plain	plain	plain
T920711.115	w/s	w/s	w/s
T920711.116	w/s	w/s	w/s
Stage 2			
T920711.117	w/s	plain	w/s
T920711.129	plain	w/s	w/s
		⋮	
T920711.213	w/s	plain	w/s
T920711.214	plain	w/s	w/s

Table 5: *Data sets given to the three annotators (plain: no suggestions, w/s: with suggestions)*

annotated earlier by two more experienced annotators (named C_{ref} and D_{ref}). As mentioned previously, C participated both in the annotation of the data for current experiment and the data for the reference annotation.

In stage 1 of the experiment we found that annotators A and B are fairly equal in their annotation skills and speed, which eased our concerns of the comparability of the results.

First of all, we wanted to see if the presence of suggestion leads to a bias in the annotation towards or against certain decisions. For this, we compared the stage 1 data, which has been annotated with the same method, respectively, with the stage 2 data, on which annotators A and B always had differing methods regarding pre-annotation (see table 6). As we had suspected, the agreement for the same-method condition is higher (both in terms of F-measure and in terms of κ) than in the different-method condition, indicating that the annotation with suggestion exhibits some bias.

On the other hand, the annotator agreement for the new annotators (even in the same-method condition) is markedly less than for the experienced annotators who created the reference annotations, so this may indicate that suggestion-based annotation and annotation from scratch lead to different types of mistakes. To check this, we compared the subset of the annotations for annotators A and B to the reference annotations. The difference between the new annotators A and B is the largest while there is less deviation of both A and B from the two reference annotations. In other words, our hypothesis is confirmed that the main cause for the difference is that pre-annotation and annotation from scratch lead to different kinds of errors rather than to a difference in annotation quality. It should therefore be possible to exploit the greater diversity in both cases by letting annotators use different annotation methods (one plain, and one with suggestions) to be able to detect more (potential) errors in a later adjudication step.

With respect to annotation times, we only found a small decrease for the pre-annotation condition ($p \approx 0.08$ using a paired t-test), whereas the difference between novice and experienced annotators is much larger and with high statistical significance ($p < 0.03$ using a paired t-test). Because of high variance of annotation time between documents, the difference in mean annotation time between pre-annotation and manual annotation seems relatively unimportant (cf. table 7).

Annotators		Method	F	κ
C_{ref}	D_{ref}	plain	0.85	0.81
A	B	stage 1	0.79	0.73
A	B	stage 2	0.76	0.67
A	C_{ref}	w/s	0.80	0.72
B	C_{ref}	w/s	0.79	0.71
A	D_{ref}	w/s	0.78	0.69
B	D_{ref}	w/s	0.77	0.69
A	C_{ref}	plain	0.79	0.74
B	C_{ref}	plain	0.79	0.73
A	D_{ref}	plain	0.78	0.73
B	D_{ref}	plain	0.78	0.72

Table 6: *Inter-annotator Agreement*

Annotator	Subset	Average	Std.dev.
A	all	874	487
B	all	957	511
C	all	385	320
A,B	plain	1042	433
A,B	w/s	726	543

Table 7: *Annotation times (in seconds)*

6 Conclusion

We investigated the use of pre-annotation for the purpose of quantifying possible improvements in speed and/or quality in the annotation of coreference. Our findings from a controlled experiment involving two novice and one experienced annotator indicate that annotation speed can be improved by a small amount, while annotator agreement with the existing reference annotation is the same (using F-measure) or only slightly worse (using κ). Our data further suggests that pre-annotation leads to different kinds of errors rather than worse annotation quality, which means that its use can be beneficial for finding more errors in the subsequent adjudication step, leading to better overall quality.

Regarding the hypotheses mentioned earlier, we find that there is a is an increase in annotation speed, even though it is relatively small (both compared to between-document variance and to the difference between novice and expert annotators), and regarding the annotation quality, it seems that neither the hope to improve annotation quality by reducing the portion of repetitive work nor the fear of worse annotation quality due to lenient annotators could be definitely confirmed, but it seems that these effects are more or less balanced. As we used an off-the-shelf annotation tool (PALinkA) with only minimal modifications in the annotation process, it is possible that tighter integration of the prediction component will lead to a larger improvement in annotation speed.

7 Acknowledgments

We would like to thank our annotators Janina Kopp, Vera Möller, Simone Müller, and Dominikus Wetzel for their dedicated work, as well as Constantin Orasan for providing advice on PALinkA and making the technical modifications we asked for even while travelling.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC MapTask Corpus. *Language and Speech*, 34(4):351–366.

Brants, T. and Plaehn, O. (2000). Interactive corpus annotation. In *LREC-2000*.

Bruneseaux, F. and Romary, L. (1997). REG: Reference Encoding Guideline. See also: Bruneseaux F., Romary L. (1997). Codage des références et coréférences dans les dialogues homme-machine. Proc. Joint Intermational Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, Kingston (Ontario).

Cahill, A., McCarthy, M., van Genabith, J., and Way, A. (2002). Parsing with PCFGs and Automatic F-Structure Annotation. In *Proceedings of the Seventh International Conference on LFG*. CSLI Publications.

Charniak, E. and Johnson, M. (2005). Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2005). TiMBL: Tilburg Memory Based Learner– version 5.1–Reference Guide. Technical Report ILK 01-04, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

Fligelstone, S. (1992). Developing a Scheme for Annotating Text to Show Anaphoric Relations. In Leitner, G., editor, *New Directions in Corpus Linguistics*, pages 153–170. Mouton de Gruyter, Berlin.

Foth, K. A., Daum, M., and Menzel, W. (2004). Interactive grammar development with WCDG. In *ACL-2004 Interactive Posters and Demonstrations*.

Hamp, B. and Feldweg, H. (1997). GermaNet – a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

Hinrichs, E., Filippova, K., and Wunsch, H. (2005). What treebanks can do for you: Rule-based and machine-learning approaches to anaphora resolution in German. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT'05)*.

Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H., and Trushkina, J. (2004). Recent developments of Linguistic Annotations of the TüBa-D/Z Treebank. In *Proceedings of TLT 2004*.

Hirschman, L. (1997). MUC-7 Coreference Task Definition, Version 3.0. In *Proceedings of MUC-7*. Science Applications International Corporation. See www.muc.saic.com.

Hockenmaier, J. and Steedman, M. (2002). Acquiring compact lexicalized grammars from a cleaner treebank. In *LREC 2002*.

Kaplan, R. M. and Maxwell, J. T. (1996). LFG grammar writer's workbench. Technical report, Xerox PARC.

Magerman, D. M. (1995). Statistical decision-tree models for parsing. In *ACL'1995*.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Miyao, Y. and Tsujii, J. (2005). Probabilistic disambiguation models for wide-coverage HPSG parsing. In *ACL 2005*.

Müller, C. and Strube, M. (2003). Multi-Level Annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 198–207, Sapporo, Japan.

Naumann, K. (2006). Manual for the Annotation of in-document Referential Relations. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

Orasan, C. (2005). PALinkA – Perspicuous and Adjustable Links Annotator. `http://clg.wlv.ac.uk/projects/PALinkA/`.

Passoneau, R. (1997). Instructions for applying discourse reference annotation for multiple applications (drama).

Passonneau, R. (1997). Applying reliability metrics to co-reference annotation. Technical Report CUCS-025-03, Columbia University.

Poesio, M. (2000). Coreference. In Mengel, A., Dybkjaer, L., Garrido, J., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C., editors, *MATE Deliverable D2.1. MATE Dialogue Annotation Guidelines*, pages 134–187.

Poesio, M. (2004). The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of SIGDIAL*, Boston.

Sampson, G. (1993). The SUSANNE corpus. *ICAME Journal*, 17:125–127.

Telljohann, H., Hinrichs, E. W., Kübler, S., and Zinsmeister, H. (2006). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z), Revised Version. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

van der Beek, L., Bouna, G., Malouf, R., and van Noord, G. (2002). The Alpino dependency treebank. In *Computational Linguistics in the Netherlands (CLIN-03)*.

Versley, Y. (2006). A constraint-based approach to noun phrase coreference resolution in German newspaper text. In *Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2006)*.

Versley, Y. (2007). Using the Web to resolve coreferent bridging in German newspaper text. In *Proceedings of GLDV-Frühjahrstagung 2007*, Tübingen. Narr.

Vieira, R. and Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*. Morgan Kaufmann.

Zinsmeister, H., Kuhn, J., and Dipper, S. (2002). TIGER Transfer: Utilizing LFG Parses for Treebank Annotations. In *LFG-02*.

Zinsmeister, H., Witt, A., Kübler, S., and Hinrichs, E. (2007). Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. To be published in: Anke Lüdeling and Merja Kytö (eds.) Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin.