A Data-driven Approach to Pronominal Anaphora Resolution for German

Erhard W. Hinrichs, Katja Filippova, Holger Wunsch
SfS-CL, University of Tübingen
Wilhelmstr. 19
72074 Tübingen, Germany
{eh,ephilpva,wunsch}@sfs.uni-tuebingen.de

Abstract

This paper reports on a hybrid architecture for computational anaphora resolution (CAR) of German that combines a rule-based pre-filtering component with a memory-based resolution module (using the Tilburg Memory Based Learner – TiMBL). The data source is provided by the TüBa-D/Z treebank of German newspaper text (Telljohann *et al.* 04) that is annotated with anaphoric relations. The CAR experiments performed on these treebank data corroborate the importance of modelling aspects of discourse structure for robust, data-driven anaphora resolution. The best result with an F-measure of 0.734 achieved by these experiments outperforms the results reported by (Schiehlen 04), the only other study of German CAR that is based on newspaper treebank data.

1 Introduction

The present study focuses exclusively on the resolution of pronominal anaphora with NP antecedents for German, where the term *pronoun* is used as a cover term for 3rd person reflexive, possessive, and personal pronouns. The purpose of this paper is threefold:

(i) to apply the machine learning paradigm of memory-based learning to the task of CAR for German,

(ii) to provide a series of experiments that corroborate the importance of modelling aspects of discourse structure for robust, data-driven anaphora resolution and that induce more fine-grained information from the data than previous approaches,

(iii) to apply CAR to a corpus of German newspaper texts, yielding competitive results for a genre that is known to be considerably more difficult than the Heidelberg corpus of tourist information texts (see (Kouchnir 03) for more discussion on this issue.)

2 Previous Research on CAR

Computational anaphora resolution has been a very active research area in computational linguistics for more than three decades. While early work on CAR was carried out almost exclusively in a rule-based paradigm, there have been numerous studies during the last ten years that have demonstrated that machine-learning and statistical approaches to CAR can offer competitive results to rule-based approaches. In particular, this more recent work has shown that the hand-tuned weights for anaphora resolution introduced by (Lappin & Leass 94), by (Kennedy & Boguraev 96), and (Mitkov 02) can be successfully simulated by data-driven methods (Preiss 02b).

While there is a rich diversity of methods that have been applied to CAR, there is also a striking convergence of grammatical features that are used as linguistic knowledge across different algorithms.[1] Most approaches base their resolution algorithm on some combination of distance between pronouns and potential antecedents, grammatical agreement between pronouns and antecedents, constituent structure information, grammatical function assignment for potential antecedents, and the type of NP involved (e.g. whether it is definite or indefinite). The combined effect of these features is to establish a notion of *discourse salience* that can help rank potential antecedents. An important aspect of discourse salience is its dynamic character since there seems to be a strong correlation between salience and discourse recency. This aspect of salience was first captured by (Lappin & Leass 94) and by (Kennedy & Boguraev 96) through the use of a decay function that decreases the score of a potential antecedent each time a new sentence is processed. In data-driven approaches this decay function is simulated by the distance measure between

[1] See (Tetreault 05) for a comprehensive survey.

pronoun and antecedent.

With the exception of the Bayesian model of (Ge *et al.* 98) and the maximum-entropy system of (Kehler 97), most data-driven approaches to CAR are based on machine learning techniques, with decision trees as the widely used paradigm (McCarthy & Lehnert 95; Soon *et al.* 01; Ng & Cardie 02; Strube & Müller 03).

Previous studies of CAR have focused on English and have been based on text corpora of fairly modest size, however see (Ge *et al.* 98) for an exception. The only previous studies for German have been presented by (Strube & Hahn 99), based on centering theory, (Müller *et al.* 02), using co-training, and by (Kouchnir 03), who applies boosting. (Schiehlen 04) provides an overview of adapting CAR algorithms to German that were originally developed for English.

While memory-based learning (MBL) has been successfully applied to a wide variety of NLP tasks, there has been only one previous study of CAR using MBL (Preiss 02a). In contrast to decision trees that have been applied to CAR by a variety of authors, memory-based learning suffers less from problems of overfitting due to its lack of data abstraction. It is also known to be more sensitive to pockets of exceptions in the data – a feature characteristic of natural language data.

3 Data

The present research focuses on German and utilizes the TüBa-D/Z (Telljohann *et al.* 04), a large treebank of German newspaper text that has been manually annotated with constituent structure and grammatical relations such as *subject, direct object, indirect object* and *modifier*. These types of syntactic information have proven crucial in previous CAR algorithms. More recently, the TüBa-D/Z annotations have been further enriched to also include anaphoric relations (Hinrichs *et al.* 04), thereby making the treebank suitable for research on CAR. German constitutes an interesting point of comparison to English since German exhibits a much richer inflectional morphology and a relatively free word order at the phrase level.

The sample sentences in (1) illustrate the annotation of referentially dependent relations in the TüBa-D/Z anaphora corpus.

(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 von
Thöne. He was yesterday with 217 out of
355 Stimmen gewählt.
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.’

In (1) 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].[2] Since noun phrases [1] and [2] are coreferential, there exists an implicit anaphoric relation between NP [1] and NP [3], with all three NPs belonging to the same coreference chain. In keeping with the MUC-6 annotation standard[3], the anaphoric relation of a pronoun is established only to its most recently mentioned antecedent. (1) also illustrates the longest-match principle for identifying markables. In case of complex NPs, the entire NP counts as a markable, but so do its subconstituents.[4] Thus, part of the CAR task consists in determining that in the case at hand the complex NP as a whole is the correct antecedent for the pronoun *er*, and not only the sub-NP *der neue Vorsitzende*.

The TüBa-D/Z currently consists of 766 newspaper texts with a total of 15260 sentences and an average number of 19.46 sentences per text. The TüBa-D/Z contains 7606 reflexive and personal pronouns, 2195 possessive pronouns, and 99585 markables (i.e. potential antecedent NPs). The number of pronouns in the TüBa-D/Z corpus is considerably larger than in the hand-annotated portion of the German NEGRA newspaper corpus (2198 possessive pronouns, 3115 personal pronouns) utilized in (Schiehlen 04) and substantially larger than the German Heidelberg tourism information corpus (36924 tokens, 2179 anaphoric NPs) used by (Müller *et al.* 02) and by (Kouchnir 03).

[2] Even though the referent of the personal pronoun [3] is the same as the referent of the noun phrases [1] and [2], the relation between a pronoun and its antecedent is taken to be anaphoric, rather than coreferent. See (vanDeemter & Kibble 00) for a detailed discussion of principled reasons not to conflate the terms *coreferent* and *anaphoric*.

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

[4] This means that in example (1) the NP *Der neue Vorsitzende* and the NP *der Gewerkschaft Erziehung und Wissenschaft* are separate markables. However, the latter will be filtered out by the XIP-module (described in section 4) since its gender (feminine) does not match the gender of the pronoun (masculine).

pronoun/antecedent	cataphoric	parallel	clause-mate	distance
discourse history	ON	OD	OA	PRED
	MOD	OPP	FOPP	APP
	TITLE	CONJ	HD	OTHER
pronoun	reflexive	possessive		

Table 1: Feature Set

4 Experiments

The experiments are based on a hybrid architecture that combines a rule-based pre-filtering module with a memory-based resolution algorithm. In the memory-based encoding used in the experiments, anaphora resolution is turned into a binary classification problem. If an anaphoric relation holds between an anaphor and an antecedent, then this is encoded as a positive instance. If no anaphoric relation holds between a pronoun and an NP, then this encoded as a negative instance.

The purpose of the pre-filtering module, which has been implemented in the Xerox Incremental Deep Parsing System (XIP) (Aït-Mokhtar *et al.* 02), is to retain only those NPs as potential antecedents that match a given pronoun in number and gender. Due to the richness of inflectional endings in German, this pre-processing step is crucial for cutting down the size of the search space of possible antecedents. Without XIP pre-filtering, the TüBa-D/Z corpus yields a total of 1,412,784 of anaphor/candidate-antecedent pairs. This number represents all possible ways of pairing a pronoun with an antecedent NP in each of the 766 texts of the TüBa-D/Z corpus. After pre-filtering this number is reduced to appr. 190,000 pairs.

The memory-based resolution module utilizes the Tilburg Memory Based Learner (TiMBL), version 5.1 (Daelemans *et al.* 05). Unless otherwise specified, the experiments use the default settings of TiMBL.

4.1 Feature Set

In the experiments, the TiMBL learner was presented with the set of features summarized in table 1. The features on line 1 all refer to relational properties of the pronoun and potential antecedents. The feature *parallel* encodes whether the anaphor and the potential antecedent have the same grammatical function. The features on line 3 refer to the pronoun alone and encode whether it is possessive or reflexive. The features on line 2 are designed to model the discourse history in terms of the grammatical functions of NPs that are in the same coreference class as the candidate antecedent. The grammatical functions are those provided by the syntactic annotation of the TüBa-D/Z treebank: ON (for: *subject*), OA (for: *direct object*), OD (for: *dative object*), PRED (for: *predicative complement*), MOD (for: *modifier*), etc.

The main purpose of the experiments reported here was to systematically study the impact that information about discourse context has on the performance of data-driven approaches to CAR. To this end, we designed two experiments that differ from each other in the amount of information about the coreference chains that are encoded in the training data.

4.2 Knowledge-Rich Encoding of Instances – Experiment I

In Experiment I, complete information about coreference chains is used for training. In example (1) the three bracketed NPs form one coreference chain since the first two NPs are coreferent and the pronoun is anaphoric to both. Accordingly, for example (1), two positive instances are created as shown in table 2. The sequence of features in each vector follows the description of features shown in table 1. Binary features are encoded as *yes*/*no*. Numeric features are given values from 1 to 30, with a special value of 31 reserved for the value *undefined*. Inspection of the data showed that a context window of this size contains the antecedent in more than 99% of all cases. For technical reasons, the numeric values are prefixed by a dash in order for TiMBL to treat them as discrete rather than continuous values. In the case at hand, the closest member of the same coreference class is in the previous sentence. Thus, the distance feature has value -1.

The first vector in table 2 displays the pairing of the pronoun with the NP *der neue Vorsitzende der Gewerkschaft Erziehung und Wissenschaft*, the first NP in the text. This NP is the subject (ON) of its clause. The value for this grammatical function is -1 since the NP occurs in the clause immediately preceding the pronoun. The second vector pairs the two preceding NPs with the pronoun *er*. Since the NP *Ulli Thöne* is in predicative position (PRED) and occurs in the same clause as the subject NP *der neue Vorsitzende der Gewerkschaft Erziehung und Wissenschaft*, the value

cat,	par,	cl-mate,	dist,	ON,	OD,	OA,	PRED,	MOD,	OPP,	FOPP,	APP,	TITLE,	CONJ,	HD,	OTHER,	refl,	poss;	class
< no,	no,	no,	-1,	-1,	-31,	-31,	-31,	-31,	-31,	-31,	-31,	-31,	-31,	-31,	-31,	no,	no;	yes >
< no,	no,	no,	-1,	-1,	-31,	-31,	-1,	-31,	-31,	-31,	-31,	-31,	-31,	-31,	-31,	no,	no;	yes >

Table 2: Sample Instances

for these two grammatical functions ON and PRED is -1. Thus, the intended semantics of the features for each grammatical function is to encode the distance of the last occurrence of a member of the same coreference class with that particular grammatical function.[5] One aspect of the discourse history that the current encoding does not model is the frequency with which a given grammatical function occurs in the text, since the encoding only registers the most recent occurrence of a given grammatical function. To control for this, a variant of the experiments reported here was conducted where for each grammatical function a pair of values was introduced consisting of the distance of the closest antecedent NP and the number of times that grammatical function appeared in the same coreference class. However, such additional mention counts did not significantly change the results of the expermiments and were therefore omitted from the feature vectors.

The sample vectors in table 2 illustrate the incremental encoding of instances. The initial vector encodes only the relation between the pronoun and the antecedent first mentioned in the text. Each subsequent instance adds one more member of the same coreference class. This incremental encoding follows the strategy of (Kennedy & Boguraev 96) and reflects a dynamic modelling of the discourse history. The last item in the vector, which is separated from the other entries by a semicolon, indicates class membership. In the memory-based encoding used in the experiments, anaphora resolution is turned into a binary classification problem. If an anaphoric relation holds between an anaphor and an antecedent, then this is encoded as a positive instance, i.e., as a vector ending in *yes*. If no anaphoric relation holds between a pronoun and an NP, then this encoded as a negative instance, i.e., as a vector ending in *no*.

4.3 Knowledge-poor Encoding of Instances – Experiment II

Experiment II uses a more knowledge-poor encoding of the data and pairs each pronoun only with the most recent antecedent in the same coreference class, thereby losing both information inherent in the entire coreference class and at the same time truncating the discourse history. Using example (1) once more as an illustration, two positive instances are created. The first vector is the same as in Experiment I. The second vector retains value -1 only for PRED, the grammatical function of the candidate itself. The value of ON is now undefined (-31).

4.4 Two Variants

For each of the two experiments described above, two variants were conducted. In one version, the evaluation focused on the closest antecedent to calculate the result for recall, precision and F-measure.[6] In a second variant, the most confident antecedent was chosen. The confidence measure was calculated by the function $\mathrm{conf}(t, c_k)$ defined as follows:

Definition Given classes $c_1 \ldots c_n$, and class distributions $d_1 \ldots d_n$ (where d_i is the number of neighbors that classified the test instance t as belonging to class c_i), the confidence $\mathrm{conf}(t, c_k)$ in the final classification c_k is

$$\mathrm{conf}(t, c_k) = \frac{d_k}{\sum_{i=1}^{n} d_i}$$

5 Evaluation

To assess the difficulty of the pronoun resolution task for the TüBa-D/Z corpus, we established as a baseline a simple heuristic that picks the closest preceding subject as the antecedent. This baseline is summarized in table 3 together with results of the experiments described in the previous section. For each experiment ten-fold cross-validation was performed, using 90% of the corpus for training and 10% for testing.

5.1 Results of Experiments I and II

Both experiments significantly outperform the baseline approach in F-measure. The findings summarized in table 3 corroborate the importance of modelling the discourse history for pronoun resolution since the results of Experiment I are consistently better than those of Experiment II. An explicit modelling of the

[5] A similar encoding is also used by (Preiss 02a).

[6] Throughout this paper the term *F-measure* implies the parameter setting of $\beta = 1$.

	av. precision	av. recall	av. F-measure
Baseline	0.500	0.647	0.564
Experiment I			
closest antecedent	0.826	0.640	0.721
most conf. antecedent	0.801	0.621	0.700
Experiment II			
closest antecedent	0.779	0.600	0.678
most conf. antecedent	0.786	0.606	0.684

Table 3: Summary of Results

6 most informative features:	clause-mate,parallel,possessive,FOPP,ON,OD
3 least informative features:	TITLE, distance,CONJ

Table 4: Summary of Feature Weights Based on GainRatioValues

discourse history with a hand-coded decay function was first proposed by (Lappin & Leass 94) and by (Kennedy & Boguraev 96). The present paper does not have to rely on the hand-coding of such a decay function. Rather, it induces the relevant aspects of the discourse history directly from the instance base used by the memory-based learner.

It is also noteworthy that in Experiment I the strategy of picking the closest antecedent outperforms the strategy of picking the most confident antecedent chosen by TiMBL.

5.2 Benchmarking Feature Impact

It is instructive to benchmark the importance of the features used in the experiments. This can be ascertained from the weights that the gain ratio measure (as the default feature weighting used by TiMBL) assigns to each feature. Gain ratio is an entropy-based measure that assigns higher weights to more informative features. Table 4 displays the top six most informative features and the three least informative features in decreasing order of informativeness. The fact that the features *clause-mate*, *parallel*, and *possessive* are the three most informative features concurs with the importance given to such features in hand-crafted algorithms for CAR. However, the ranking of some of the features included in table 4 is rather unexpected. The fact that the grammatical function FOPP (for: *optional PP complement*) outranks the grammatical function *subject* (ON) runs counter to hand-coded salience rankings found in the literature which give the feature *subject* the highest weights among all grammatical functions. That the FOPP feature outranks the function *subject* is due to the fact that the presence of an optional PP-complement is almost exclusively paired with negative instances. This finding points to an important advantage of data-driven approaches over hand-crafted models. While the latter only take into account positive evidence, data-driven models can profit from considering positive and negative evidence alike. Perhaps the most surprising result is the fact that *distance* between anaphor and antecedent is given the second lowest weight among all eighteen features. This sharply contrasts with the intuition often cited in hand-crafted approaches that the distance between anaphor and antecedent is a very important feature for an adequate resolution algorithm. The reason why distance receives such a low weight might well have to do with the fact that this feature becomes almost redundant when used together with the other distance-based features for grammatical functions.

The empirical findings concerning feature weights summarized in table 4 underscore the limitation of hand-crafted approaches that are based on the analysts' intuitions about the task domain. In many cases, the relative weights of features assigned by data-driven approaches will coincide with the weights assigned by human analysts and fine-tuned by trial and error. However, in some cases, feature weightings obtained automatically by data-driven methods will be more objective and diverge considerably from manual methods, as the weight assigned by TiMBL to the feature *distance* illustrates.

5.3 Optimization by Fine-tuning of TiMBL Parameters

It has been frequently observed (e.g. by (Hoste *et*

	av. precision	av. recall	av. F-measure
Baseline	0.500	0.647	0.564
Experiment I closest antecedent	**0.827**	**0.661**	**0.734**

Table 5: Summary of Best Results

al. 02)) that the default settings provided by a classifier often do not yield the optimal results for a given task. The CAR task for German is no exception in this regard. TiMBL offers a rich suite of parameter settings that can be explored for optimizing the results obtained by its default settings. Some key parameters concern the choice of feature distance metrics, the value of k for the number of nearest neighbors that are considered during classification as well as the choice of voting method among the k-nearest neighbors used in classification. TiMBL's default settings provide the feature distance metric of weighted overlap (with the gain ratio measure for feature weighting), $k = 1$ as the number of k-nearest neighbors, and majority class voting.

To assess the possibilities of optimizing the results of Experiments I and II, the best result (Experiment I with closest antecedent) was chosen as a starting point. The best results, shown in table 5, were obtained by using TiMBL with the following parameters: modified value distance metric (MVDM), no feature weighting, $k = 3$, and inverse distance weighting for class voting.

The optimizing effect of the parameters is not entirely surprising.[7] The MVDM metric determines the similarities of feature values by computing the difference of the conditional distribution of the target classes for these values.[8] For informative features, $\delta(v_1, v_2)$ will on average be large, while for less informative features will tend to be small. (Daelemans *et al.* 05) report that for NLP tasks MVDM should be combined with values of k larger than one. The present task confirms this result by achieving optimal results for a value of $k = 3$.

6 Comparison with Related Work

The only previous study of German CAR that is based on newspaper treebank data is that of (Schiehlen 04).[9] Schiehlen compares an impressive collection of published algorithms, ranging from reimplementations of rule-based algorithms to reimplementations of machine-learning and statistical approaches. The best results of testing on the NEGRA corpus were achieved with an F-measure of 0.711 by a decision-tree classifier, using C4.5 and a pre-filtering module similar to the one used here. The best result with an F-measure of 0.734 achieved by the memory-based classifier and the XIP-based pre-filtering component outperforms Schiehlen's results, although a direct comparison is not possible due to the different data sets.

7 Summary and Future Work

The current paper presents a hybrid architecture for computational anaphora resolution (CAR) of German that combines a rule-based pre-filtering component with a memory-based resolution module (using the Tilburg Memory Based Learner – TiMBL). The data source is provided by the TüBa-D/Z treebank of German newspaper text that is annotated with anaphoric relations. The CAR experiments performed on these treebank data corroborate the importance of modelling aspects of discourse structure for robust, data-driven anaphora resolution. The best result with an F-measure of 0.734 achieved by the memory-based classifier and the XIP-based pre-filtering component outperforms Schiehlen's results, although a direct comparison is not possible due to the different data sets.

The experiments reported here are all based on treebank data. In future work it is planned to use the output of a robust parser for German as input to the hybrid model presented here. Several parsers are good candidates for such an extension. The parsers for German developed by (Trushkina 04), (Müller 05) and by (Foth *et al.* 04) all produce the relevant grammatical

[7] See (Hoste *et al.* 02) for the optimizing effect of MVDM in the word sense disambiguation task.

[8] More specifically, the distance $\delta(v_1, v_2)$ between two feature values v_1 and v_2 is defined as

$$\delta(v_1, v_2) = \sum_{i=1}^{n} \mid P(C_i|v_1) - P(C_i|v_2) \mid$$

[9] (Kouchnir 03) briefly discusses results of applying her ensemble learning classifier to a hand-annotated corpus of the German weekly newspaper *Der Spiegel*. However, compared to her results on the Heidelberg tourism corpus, the best results for the *Spiegel* data are rather low with an F-measure of 34.4 %.

information needed for the features employed by the memory-based module.

References

(Aït-Mokhtar *et al.* 02) Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2–3):121–144, 2002.

(Daelemans *et al.* 05) Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. TiMBL: Tilburg memory based learner– version 5.1–reference guide. Technical Report ILK 01-04, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University, 2005.

(Foth *et al.* 04) Kilian Foth, Michael Daum, and Wolfgang Menzel. A broad-coverage parser for german based on defeasible constraint. In *KONVENS 2004, Beiträge zur 7. Konferenz zur Verarbeitung natürlicher Sprache*, pages 45–52, Vienna, 2004.

(Ge *et al.* 98) Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, Montreal, Canada, 1998.

(Hinrichs *et al.* 04) Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In Sandra Kübler, Joakim Nivre, Erhard Hinrichs, and Holger Wunsch, editors, *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, Tübingen, Germany, 2004.

(Hoste *et al.* 02) Veronique Hoste, Iris Hendrickx, Walter Daelemans, and Antal van den Bosch. Parameter optimization for machine-learning of word sense disambiguation. *Natural Language Engineering*, 8(4):311–325, 2002.

(Kehler 97) Andrew Kehler. Probabilistic coreference in information extraction. In *Proceedings of the Second Conference on Empirical Methods in NLP (EMNLP-2)*, Providence, RI, 1997.

(Kennedy & Boguraev 96) Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *The Proceedings of the 16th International Conference on Computational Linguistics*, 1996.

(Kouchnir 03) Beata Kouchnir. A machine learning approach to German pronoun resolution. Unpublished M.Sc. thesis, School of Informatics, University of Edinburgh, 2003.

(Lappin & Leass 94) Shalom Lappin and Herbert Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

(McCarthy & Lehnert 95) Joseph F. McCarthy and Wendy G. Lehnert. Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1050–1055, Montreal, Canada, 1995.

(Mitkov 02) Ruslan Mitkov. *Anaphora Resolution*. John Benjamins, Amsterdam, 2002.

(Müller 05) Frank H. Müller. *'A Finite-State Approach to Shallow Parsing and Grammatical Functions Annotation of German*. Unpublished PhD thesis, University of Tübingen, 2005.

(Müller *et al.* 02) Christoph Müller, Stefan Rapp, and Michael Strube. Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 352–359, Philadelphia, PA, USA, 2002.

(Ng & Cardie 02) Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 104–111, Philadelphia, PA, USA, 2002.

(Preiss 02a) Judita Preiss. Anaphora resolution with memory based learning. In *Proceedings of the 5th UK Special Interest Group for Computational Linguistics (CLUK5)*, pages 1–8, 2002.

(Preiss 02b) Judita Preiss. A comparison of probabilistic and non-probabilistic anaphora resolution algorithms. In *Proceedings of the Student Workshop at the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 42–47, Philadelphia, PA, USA, 2002.

(Schiehlen 04) Michael Schiehlen. Optimizing algorithms for pronoun resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, August 2004.

(Soon *et al.* 01) Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

(Strube & Hahn 99) Michael Strube and Udo Hahn. Functional centering - grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344, 1999.

(Strube & Müller 03) Michael Strube and Christoph Müller. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL-03*, pages 168–175, Sapporo, July 2003.

(Telljohann *et al.* 04) Heike Telljohann, Erhard W. Hinrichs, and Sandra Kübler. The TüBa-D/Z Treebank – Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 2004.

(Tetreault 05) Joel Tetreault. *Empirical Evaluations of Pronoun Resolution*. Unpublished PhD thesis, University of Rochester, Department of Computer Science, 2005.

(Trushkina 04) Julia S. Trushkina. *Morpho-syntactic Annotation and Dependency Parsing of German*. Unpublished PhD thesis, University of Tübingen, 2004.

(vanDeemter & Kibble 00) Kees van Deemter and Roger Kibble. On coreferring: Coreference annotation in muc and related schemes. *Computational Linguistics*, 26(4):615–623, 2000.