What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German.

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

This paper compares two approaches to computational anaphora resolution for German: (i) an adaption of the rule-based RAP algorithm that was originally developed for English by Lappin and Leass, and (ii) a hybrid system for anaphora resolution that combines a rule-based pre-filtering component with a memory-based resolution module. The data source is provided by the TüBa-D/Z treebank of German newspaper text (Telljohann et al., 2003) that is annotated with anaphoric relations. This treebank uses as its data source a collection of articles of the German daily newspaper *taz* (*die tageszeitung*).

Due to their fine grained syntactic annotation, the TüBa-D/Z treebank data are ideally suited as a basis for the identification of markables for pronominal reference 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 order regularities among different clause types of German and which are widely accepted among descriptive linguists of German (cf. e.g. 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 (1). 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:

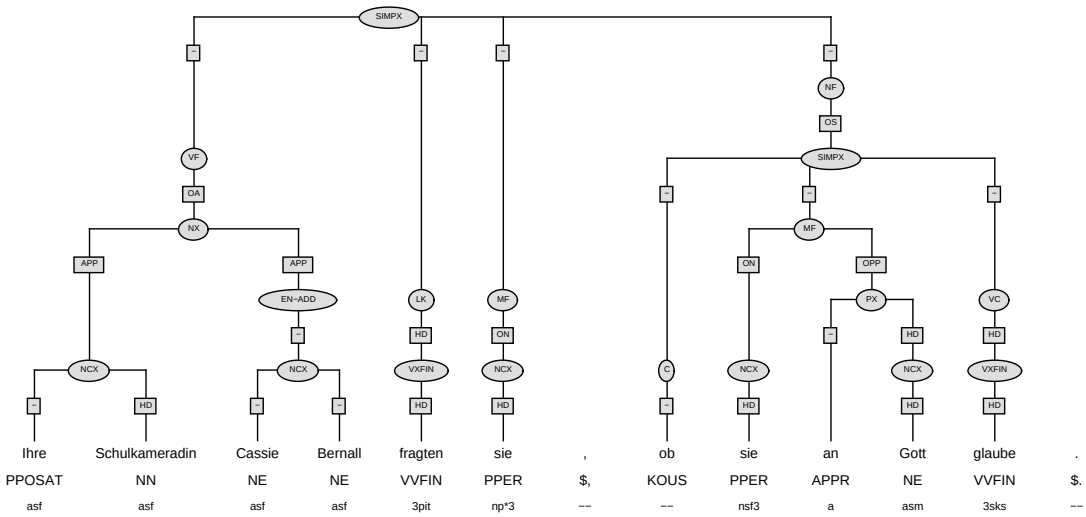

Figure 1: A sample tree from the TüBa/D-Z treebank.

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*).

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

Topological field information and grammatical function information is important 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) or Lappin and Leass (1994), rely on such binding-constraint filters to minimize the set of potential antecedents for pronouns and reflexives.

1 Rule-based anaphora resolution for German

The Resolution of Anaphora Procedure (RAP; Lappin and Leass 1994) relies on measures of salience derived from syntactic structure and a dynamic model of attentional state to select the NP antecedent of a third person pronoun. For the present paper, a German version of RAP (RAP-G) has been implemented. RAP-G employs both a morphological and a syntactic filter to reduce the number of candidate pronoun-antecedent pairs that are passed on to the resolution module.

1.1 The morphological filter

Unlike the original English version, which uses a built-in morphological filter, RAP-G delegates the morphological filtering process to an external module, which has been implemented in the Xerox Incremental Deep Parsing System (XIP; Aït-Mokhtar et al., 2002). The purpose of the morphological filter is to retain only those NPs as potential antecedents that match a given pronoun in number and gender. Given the richness of inflectional endings in German, this pre-processing step is highly effective in cutting down the size of the search space of possible antecedents.

1.2 The syntactic filter

The pronoun-NP pairs that have been admitted by the morphological filter are passed on to the syntactic filter. Only pairs of a personal or reflexive pronoun and a candidate antecedent that do not violate the constraints of binding-theory pass this filter. The resulting set of candidates, which meets both morphological and syntactic constraints, serves as the input to the actual resolution module.

Factor type	Weight
Subject emphasis	170
Accusative object emphasis	70
Dative object emphasis	50
Genitive object emphasis	50
Head noun emphasis	80

Table 1: Grammatical role hierarchy used by RAP-G

1.3 The resolution module

The resolution module is the central component in RAP-G. For each pronoun to be resolved, RAP-G selects an antecedent from the filtered list of candidates. The resolution module assumes for each NP an associated discourse referent, which is assigned several salience factors. Each salience factor corresponds to one of the features considered by the algorithm. Features are weighted: the more important the feature, the higher the weight. The final salience value of the discourse referent is calculated by summing up the individual weights.

RAP-G determines the salience weights according to a ranked hierarchy of grammatical roles and a model of the dynamic nature of discourse. The hierarchy is listed in table 1. The set of grammatical functions considered was adapted to the German data by splitting the single feature for "indirect object" in the original English version into two features representing the dative and genitive object. The salience weights were determined and optimized empirically. It is noteworthy that the salience weight of the grammatical function subject relative to other grammatical functions turns out to be much higher for German than for English, where Lappin and Leass empirically arrived at a salience value of 80.

The resolution proceeds sentence by sentence, from left to right. Discourse referents are first assigned salience values based on the syntactic properties of their corresponding NPs and their membership in an equivalence class (see below). The salience values are then updated on a pronoun-by-pronoun basis. Parallelism of grammatical functions between the pronoun and its potential antecedent is rewarded by increasing, potential cataphoric relations are penalized by decreasing the salience weights of the corresponding discourse referents. Unlike the English version, which strongly penalizes any potential cataphoric relationship, RAP-G distinguishes between "local" cataphora (postcedents in the same sentence) and "non-local" cataphora, where the former is less penalized than the latter.

With the following three strategies, RAP-G models the dynamic character of discourse: Both potential antecedents and postcedents receive an additional reward if they occur in the same sentence as the pronoun. It is worth noting that the optimal

Factor type	Weight
Short distance cataphora penalty	-80
Long distance cataphora penalty	-175
Parallelism reward	35
Current sentence reward	20

Table 2: Penalties and rewards used by RAP-G

weight for this reward turns out to be much lower for German than for English (20 vs. 100, see table 2).

To reflect salience decay, salience values are decreased exponentially in the distance of the sentence where the potential antecedent occurs to the current sentence according to the formula $\hat{sv} = \frac{sv}{2^{sd}}$ where sv is the prior salience value, sd the sentence distance, and $\hat{sv}$ the updated salience value.

When a pronoun is resolved to an antecedent, the discourse referents of the pronoun and the antecedent are merged into a single equivalence class. A discourse referent that is member of an equivalence class gets assigned as its salience value the sum of the salience values of all class members. This way equivalence classes mirror the effect that the salience of a discourse referent increases with its mention count.

As the final step, the resolution module selects as the antecedent or postcedent of the pronoun to be resolved the nearest NP that corresponds to the discourse referent with the highest salience value.

1.4 Results

For the experiments, all 766 articles of the TüBa-D/Z were used. They contain 1 504 possessive pronouns, 1 544 reflexive pronouns, and 2 492 personal pronouns, where only third person pronouns were considered. First person and second person pronouns were filtered out. This amounts to a total of 5 540 pronouns to be resolved. We use standard precision and recall to measure RAP-G's performance. A pair of a pronoun and an antecedent or postcedent is considered correct if they belong to the same coreference set in the manually annotated gold data. In some cases, RAP-G picks as the antecedent of a pronoun an NP that either contains the NP annotated as the correct antecedent in the gold data, or an NP that is contained in the gold antecedent. This is treated as a special case in the evaluation algorithm; such a pair is counted as correct if the contained NP is the head of the containing NP. Evaluated this way, the precision achieved by RAP-G is 76.64%, recall is 76.48%, which amounts to an F-measure of 75.56%. The results are summarized in table 3.

Possessive pronouns	1 504
Reflexive pronouns	1 534
Personal pronouns	2 490
Total pronouns	5 540
Precision	76.64%
Recall	76.48%
F-measure	76.56%

Table 3: Results of RAP-G

2 Memory-based anaphora resolution for German

As an alternative to the reimplementation of the RAP algorithm for German, we implemented a hybrid architecture that combines a rule-based pre-filtering module with a memory-based (MB) resolution algorithm. In the MB encoding used in the experiments, anaphora resolution is turned into a binary classification problem. If an anaphoric relation holds between an anaphor and an NP candidate, 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[1].

2.1 Morphological and Syntactic Filtering

Morphological filtering is achieved in the same way as in the RAP-G approach: The rule-based module filters out candidates that do not match a given pronoun in number or in gender.

The syntactic filter is not as powerful as the one used by RAP-G because it does not put any syntactic constraints on the antecedent of a reflexive. It sorts out candidates that do not satisfy the following constraints:

- A pronoun must not be contained in its antecedent
- A non-reflexive pronoun must not be contained in the antecedent's argument domain

An NP_1 is in the argument domain of an NP_2 if both NPs are arguments of the same head. Analogously, an NP_1 is in the adjuct domain of an NP_2 if NP_2 is an argument of a head and NP_1 is contained in an adjuct of the same head.

[1]Note, that any element of the coreference chain of the pronoun is considered as the right antecedent.

	Feature	Value		
1	Pronoun type	personal	possessive	reflexive
2	Position	anaphoric	cataphoric	
3	Syntactic parallelism	parallel	different	non-applicable
4	Distance in sentences	loc	0 / 1 / 2 / 3	
5	Distance in words	1..n		

Table 4: TiMBL input features

Apart from that, corpora investigation showed that only in 167 out of about 19,000 (less than 1%) cases a correct antecedent is not located within the three previous sentences. Given this, it was decided to consider a noun phrase as a candidate if it is either in the same sentence as a pronoun in consideration, or not further than three sentences before it. Concerning reflexives, the antecedent must be located in the same sentence as the pronoun.

The candidates which have passed both filters and the distance constraint serve as the input to the MB resolution module.

2.2 The memory-based resolution module

The aim of the MB implementation was to show that the performance of machine learning approaches for the anaphora resolution task is comparable to the performance of hand-crafted resolvers, if the two approaches are supplied with the same information[2]. As with the rule-based implementation, the main focus is on modelling discourse salience, specifically on finding the encoding which best captures the notion of salience.

The MB resolution module utilizes the Tilburg Memory Based Learner (TiMBL), version 5.1 (Daelemans et al., 2005). During the resolution phase TiMBL stores all the training examples in memory and for every testing instance finds the k most similar training instances. Their classes help TiMBL to assign a class to the new instance. In the case at hand, TiMBL faces a binary classification problem where *yes* stands for the anaphoric relation and *no* for its absence.

2.2.1 Different versions of the experiment

The first five features TiMBL learns from are summarised in Table 4. The very first attribute refers to the *pronoun's type*: personal, possessive, or reflexive. The next four features describe relations between the pair according to whether the pronoun

[2]Note, that Preiss (2002) reports similar results for English.

	Notation	Description
6	ON	subject
7	OA	direct object
8	OD	dative object
9	OPP	obligatory prepositional object
10	APP	apprositon
11	FOPP	optional prepositional object
12	–	a noun phrase from the title
13	X-MOD	all kinds of modifiers
14	PRED	predicative
15	KONJ	conjunct
16	HD	head of a phrase
17	-	non-head constituent of a phrase

Table 5: TüBa-D/Z syntactic functions for NPs

precedes or follows the candidate (*position*), whether they have the same syntactic function (*parallelism*[3]), and how far they are from each other in sentences and in words. If the two are located in the same clause, the numeric distance in sentences is replaced with the *loc* value. If they are from the same sentence but not from the same clause, the value is *0*. Distance in words, a feature which has proven to be helpful for other resolution algorithms (Müller et al., 2002), may have any positive number as its value[4].

Since the focus of this approach is on the salience of entities, the encoding should reflect how prominent a given entity is for the speaker, so that its further mention is pronominalised. Salience is reconstructed by the preceding mentioning of the entity. Moreover, it makes sense not only to count how often a given entity has been mentioned in the preceding text but also to see how these mentions are distributed among different syntactic functions. The set of possible syntactic functions of a noun phrase in TüBa-D/Z is presented in Table 5.

The previous mentions of a candidate refer to all markables which are coreferent with the candidate (*co-members*), as long as they occur within a specified window of sentences adjacent to the pronoun. The size of this window was determined empirically: It turned out that a window of seven sentences yielded the best results.

[3]This feature is not applicable if either the anaphor or the candidate in consideration are possessive pronouns.

[4]Since absolute distance is computed, the values are positive for both, cataphoric and anaphoric relations.

Note, that if there is more than one markable from the same set within the three-sentence window they are all considered candidates. The input instances for them may differ only in values of the following parameters: *position, parallelism, distance in words* and *sentences*. It can also be the case that the only difference between two such instances is *distance in words*. Taking only one representative from each coreferential chain would considerably decrease the number of positive training instances and would make the learning and resolving tasks much harder.

For the first version of the experiment, labeled **Experiment 1**, it was decided to simply mark how often the entity was expressed by each of the twelve syntactic roles. A TiMBL input vector looks the way given in Table 6. The attributes are set in the same order as attributes from Table 4 concatenated with the ones from Table 5: two *1*'s following *21* in the third line mean that the entity represented by the candidate in consideration has been once a subject and once an accusative object.

pers	ana	diff	2	48	0	0	0	0	0	0	0	0	0	0	1	no
pers	cat	diff	0	42	0	0	0	0	0	0	0	0	0	0	1	no
pers	ana	diff	1	21	1	1	0	0	0	0	0	0	0	0	0	yes

Table 6: Three input lines of Experiment 1

pers	ana	diff	2	48	n	n	n	n	n	n	n	n	n	n	-2	no
pers	cat	diff	0	42	n	n	n	n	n	n	n	n	n	n	-2	no
pers	ana	diff	1	21	-1	-2	n	n	n	n	n	n	n	n	n	yes

Table 7: Three input lines of Experiment 2

The second version of the experiment, labeled **Experiment 2**, does not emphasise the frequency of each mention but encodes how far from the pronoun the last mention of each syntactic function has occurred. An example of the input vector is given in Table 7 (note that *n* stands for 'never')

Having processed all the testing instances TiMBL outputs a file that looks the same way as the testing file but in the end of each line a class assigned by TiMBL is added, in our case it is either *yes* or *no*. Since TiMBL considers each candidate separately, it can not be aware if it has seen any other candidate for a pronoun or not. Being unaware of how many antecedents it has already found, if any, TiMBL does not necessarily provide exactly one antecedent. It may also be that for the same pronoun TiMBL positively resolves instances that represent members of different coreferential chains. In case of several positive instances for the same pronoun it

is decided to pick the one which is the closest to the anaphor, that is the candidate with the smallest *distance in words* value, as the ultimate antecedent.

Still, there are cases when for a given pronoun TiMBL does not find any instance which it could classify as positive, and in this case the pronoun stays unresolved. Since this is highly undesirable, an additional operation can be done then: for the unresolved pronouns the closest subject which passes the morphological and syntactic filters and which is located within the three-sentence window is picked as the antecedent. With this modification two more experiments were done, the extended **Experiment 1.subj** and the extended **Experiment 2.subj**.

So far, different encodings for discourse salience have been considered. To show that discourse information matters, an additional experiment was done, where each candidate is considered as it is. Here, the salience of the candidate is described by its grammatical role and location in respect to the anaphor. Such encoding does not fully reflect the prominence of the entity since it ignores the contribution of other elements from the same coreferential chain. This experiment is labeled **Experiment 0**, and an example of its input vector is presented in Table 8. The first five factors are the same as for Experiments 1 and 2, and the sixth factor is the syntactic function of the candidate.

pers	ana	diff	2	48	-	no
pers	cat	diff	0	42	-	no
pers	ana	diff	2	21	OA	yes

Table 8: Three input lines of Experiment 0

2.3 Results

The experiments were evaluated on the same data as RAP-G. The results of each version of the MB experiment described above are presented in Table 9.

	Precision	Recall	F-Measure
EXP 0	78,8%	63,7%	70,4%
EXP 1	83,8%	66,8%	74,3%
EXP 2	84,2%	66,4%	74,2%
EXP 1.subj	79,1%	75,1%	77%
EXP 2.subj	78,2%	74,1%	76,1%

Table 9: Results of the algorithm

The low performance of Experiment 0 compared to the results of the other experiments supports the importance of capturing discourse salience, since it is the lack of discourse information that distinguishes Experiment 0 from Experiments 1 and 2. The distance encoding of salience used in Experiment 2 turned out to be a little bit less effective than the mention counts encoding. The strategy of taking the closest subject in case of unresolved pronouns significantly increases recall but inevitably causes loss in precision. The total increase of about 2-3% in f-measure shows that this heuristics works well given that morphological and at least some syntactic prefiltering has been done.

2.3.1 Feature Ranking

For Experiment 1 the most informative features are *OD, ON, OA, parallelism, type of pronoun*, and *non-head constituency*. For Experiment 0 the three most informative features are *parallelism, syntactic function*, and *type of pronoun*. This ranking correlates with the relative feature weights of RAP-G, which also gives preference to subject, direct and dative objects and which also rewards syntactic parallelism.

3 Conclusion

This study compares two different approaches to anaphora resolution: a rule-based system employing a re-implementation of Lappin and Leass' "Resolution of Anaphora Procedure" (Lappin and Leass, 1994) that has been adapted for German, and a hybrid model combining a rule-based morphological filter with a memory-based resolution module that has been implemented using the Tilburg Memory-Based Learner (TiMBL; Daelemans et al., 2005). Both systems achieve roughly equal results, with the memory-based system (F-measure 77%) slightly outperforming the rule-based approach (F-measure 76.56%). The similarity of these results is remarkable, considering that the architecture of both systems differs fundamentally. This shows that anaphora resolution systems based on machine-learning approaches can successfully simulate the functionality of a rule-based system by automatically extracting the necessary information from the features it is presented without requiring the human effort of hand-crafted rules.

References

Aït-Mokhtar, S., J.-P. Chanod, and C. Roux (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering 8*(2–3), 121–144.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Daelemans, W., J. Zavrel, K. van der Sloot, and A. van den Bosch (2005). TiMBL: Tilburg memory based learner– version 5.1–reference guide. Technical Report ILK 01-04, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua 44*, 311–338.

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

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

Müller, C., S. Rapp, and M. Strube (2002). Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA, USA, pp. 352–359.

Pollard, C. and I. Sag (1994). *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. Chicago, IL: University of Chicago Press.

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

Telljohann, H., E. W. Hinrichs, and S. Kübler (2003). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.