# Towards Bridging Resolution in German: Data Analysis and Rule-based Experiments

**Janis Pagel and Ina Rösiger**
Institute for Natural Language Processing
University of Stuttgart, Germany
{pageljs,roesigia}@ims.uni-stuttgart.de

## Abstract

Bridging resolution is the task of recognising bridging anaphors and linking them to their antecedents. While there is some work on bridging resolution for English, there is only little work for German. We present two datasets which contain bridging annotations, namely DIRNDL and GRAIN, and compare the performance of a rule-based system with a simple baseline approach on these two corpora. The performance for full bridging resolution ranges between an F1 score of 13.6% for DIRNDL and 11.8% for GRAIN. An analysis using oracle lists suggests that the system could, to a certain extent, benefit from ranking and re-ranking antecedent candidates. Furthermore, we investigate the importance of single features and show that the features used in our work seem promising for future bridging resolution approaches.

## 1 Introduction

*Bridging* (Clark, 1975) or *associative anaphora* (Hawkins, 1978) is an anaphoric phenomenon, where a discourse-new entity stands in a prototypical or inferable relationship to a previously introduced entity. Crucially, these two entities are not coreferent.

(1) Und man muss jetzt aufpassen, dass man sich nicht zum Sprachrohr von Leuten macht, die eben den Mindestlohn umgehen wollen. Einer **der Hauptstreitpunkte** ist ja **die Dokumentationspflicht**[1].
(And now you have to be careful that you do not become the voice for the people who just want to avoid the minimum wage. One of **the main points of contention** is **the documentation requirement**...)

[1] Anaphors are marked in bold face, their antecedents are underlined.

Bridging anaphors can be considered expressions with an implicit argument, e.g. *die Dokumentationspflicht beim Mindestlohn (the documentation requirement relevant to the minimum wage)*.

The related NLP task of bridging resolution is to identify bridging anaphors and link them to their antecedents. Most of the work on bridging resolution, with its subtasks of anaphor detection and antecedent selection, has focused on English (e.g. Hou et al., 2014; Markert et al., 2012; Rahman and Ng, 2012). For German, Grishina (2016) has presented a corpus of 432 bridging pairs as well as an in-depth analysis on some properties of bridging, e.g. on the distance between anaphors and their antecedents and on the distribution of bridging relations. Apart from Cahill and Riester (2012)'s work on bridging anaphor detection as a subclass in information status classification and Hahn et al. (1996)'s early work on bridging resolution, there have been no automatic approaches to bridging resolution in German.

This paper gives an overview on German corpora containing bridging annotations and presents experiments on bridging anaphor detection and full bridging resolution on two available corpora, DIRNDL and GRAIN. The performance for full bridging resolution ranges between an F1 score of 13.6% for DIRNDL and 11.8% for GRAIN. We investigate this difference in performance by using oracle lists, which evaluate the antecedent search techniques of the rules.

## 2 Related work

### 2.1 Available corpora

This section briefly presents the three German corpora that contain bridging annotations.

**GRAIN** Recently, the GRAIN release of the SFB732 Silver Standard Collection (Schweitzer et al., 2018) has been announced. It contains



*Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 50–60
New Orleans, Louisiana, June 6, 2018. 

23 German radio interviews of about 10 minutes each, whose transcripts were annotated with referential information status (Baumann and Riester, 2012), following the annotation guidelines in Riester and Baumann (2017). This means that all referring expressions in the interviews were categorised as to whether they are given/coreferential, bridging anaphors, deictic, discourse-new, idiomatic, etc. The interviews also contain coreference chains and bridging links. 274 bridging pairs were annotated in total[2]. While the referential information status was hand-annotated, the other annotation layers consist of predicted annotations. GRAIN contains spontaneous speech about rather diverse topics.

**DIRNDL** The DIRNDL corpus (Eckart et al., 2012; Björkelund et al., 2014), a corpus of radio news, also contains bridging annotations as part of its information status annotation (again, on transcripts of the news), following older guidelines of the RefLex scheme (Baumann and Riester, 2012). Overall, 655 bridging pairs have been annotated. Apart from the manual information status annotation, other linguistic annotation layers (POS-tagging, parsing, morphological information) have been created automatically.

**Corefpro corpus** The corefpro corpus (Grishina, 2016) contains news and narrative text as well as medicine instruction leaflets, and comprises 432 annotated bridging pairs. There are three different types of anaphors: coreferent, bridging or near-identity, following Recasens and Hovy (2010). Only definite anaphors were annotated. The corpus was not available when we performed our experiments, but has recently been made publicly available[3].

### 2.2 Computational approaches

As mentioned in the introduction, there has only been little work on bridging for German so far. Cahill and Riester (2012) presented a CRF-based automatic classification of information status, which included bridging as a subclass. However, they did not state the accuracy per class, which is why we cannot derive any performance estimation for the task of bridging anaphor detection. They stated that bridging cases "are difficult to capture by automatic techniques", which confirms findings from information status classification for English, where bridging is typically a category with rather low accuracy (Markert et al., 2012; Rahman and Ng, 2012; Hou, 2016a). Hahn et al. (1996) and Markert et al. (1996) have presented a resolver for bridging anaphors, back then called textual ellipsis or functional anaphora, in which they resolved bridging anaphors in German technical texts using centering theory and a knowledge base. The corpus and the knowledge base as well as the overall system are, however, not available, which makes a comparison with our system difficult. As far as we know, the rule-based system from Hou et al. (2014) is the only system proposed for full bridging resolution so far, following earlier work on bridging anaphor detection (Hou et al., 2013a) and antecedent selection (Hou et al., 2013b).

## 3 Bridging definition in RefLex

As both available corpora, DIRNDL and GRAIN, were annotated according to the RefLex scheme (Baumann and Riester, 2012; Riester and Baumann, 2017), we present the main idea of this scheme, as well as its implications for bridging anaphors.

RefLex (Riester and Baumann, 2017) distinguishes information status at two different dimensions, namely a referential and a lexical dimension. The referential level analyses the information status of referring expressions (i.e. noun phrases) according to a fine-grained version of the given/new-distinction, whereas the lexical level analyses the information status at the word level, where content words are analysed as to whether the lemma or a related word has occurred before.

Bridging anaphors are a subclass of referential information status and are labeled as `r-bridging`. On the referential level, indefinite expressions are considered to be discourse-new and are thus treated as expressions of the information status category `r-new`. Therefore, the bridging anaphors in our data are always definite.

In RefLex, `r-briding-contained` is a separate information status class, where the anaphor is modified by the antecedent in either a prepositional modification or a possessive premodification, e.g. in *the approach's* ***accuracy*** or ***the accuracy*** *of the approach*. In this paper, we do not cover these cases.

[2] In a preliminary version of the data, in which one interview is missing as it is currently being validated.

[3] https://github.com/yuliagrishina/corefpro

## 4 Analysis: Bridging in GRAIN

Before resolving bridging references in an automatic approach, we analysed the newest of the available corpora, the GRAIN corpus, with respect to the bridging annotations, in order to get a better feeling for the annotations. As GRAIN contains natural discourse in the form of radio interviews, we believe that it is well-suited for this type of analysis.

We categorise the occurrences of bridging in GRAIN into three main categories: *prototypical*, *world-knowledge-dependent* and *unspecified*. These types reflect our intuition about the bridging phenomena in GRAIN[4]. Prototypical bridging means that the anaphor stands in a prototypical relationship to its antecedent, see Example (2). Here, *caretakers* and *patients* are prototypical members of a retirement home.

(2) Aber jetzt zum Beispiel am Bürokratiewahnsinn in den Heimen, der **den Pflegekräften** die Zeit für **die Patienten** nimmt, ändert sich ja dadurch erst einmal nichts.
(But for now, it changes nothing about the bureaucracy madness in the retirement homes, which takes all the time that **the caretakers** could spend on **the patients**.)

Prototypical relations can also be sub-categorised, leading to sub-types that others have also observed, e.g. *building-part* or *professional-role* (cf. Hou et al. (2014)). Additionally, due to GRAIN's domain, many prototypical bridging pairs are related to countries and properties of countries (see Rule 9 in Section 7.2.1).

Example (3) presents a case of bridging where world-knowledge is necessary in order to infer that *athletes* are the athletes of the sports events in Sochi for the Winter Olympics in 2014.

(3) [...], dass ich nicht nach Sotschi fahren konnte, obwohl ich als Sportlerin da wirklich sehr, sehr gerne jetzt auch in der neuen Rolle hingefahren wäre, um **die Sportler** zu unterstützen.
([...], that I couldn't go to Sochi, even though I really, really would have liked to go as an athlete and also in my new role, in order to support **the athletes**.)

Finally, many bridging anaphors do not fall in any of the other two categories, see Example (4). *Beginning* is not prototypically related to *reform* and there is no world-knowledge involved in knowing that a reform can have a beginning (it is probably more of an inference that a reform is a process, which typically has an end and a beginning).

(4) Das ist das größte Reformwerk seit Jahrzehnten in Deutschland. Und kein Wunder, dass es da **am Anfang** ruckelt.
(This is the biggest reform in Germany for decades. No wonder that it is unstable **in the beginning**.)

We manually counted the types of bridging in GRAIN and observe counts for our three main types and for the types proposed in Hou et al. (2014), as shown in Table 1. We also find instances of comparative anaphora (see Markert et al., 2012).

| Type | Sub-type | Count |
|---|---|---|
| Prototypical | Building-part | 3 |
| | Professional role | 1 |
| | Country-related | 19 |
| | Other prototypical | 69 |
| World-Knowledge | | 23 |
| Unspecified | | 101 |
| Comparative | | 8 |

Table 1: Types of bridging in GRAIN and their counts.

## 5 Experimental setup

### 5.1 Data

**GRAIN** GRAIN (Schweitzer et al., 2018) will be released soon[5]. As the annotation project is associated with our project, we have received an early version of the data, in which one of the 23 interviews is missing[6]. As no train-test-development split has yet been specified, we split the data ourselves[7].

[4] As the categorisation was performed by only one person (the first author), it has to be taken with a grain of salt. Still, we believe it is helpful to get a better feeling for the data.

[5] The release, as well as a detailed documentation is published in the framework of CLARIN 8 and available via a persistent identifier: `http://hdl.handle.net/11022/1007-0000-0007-C632-1`.

[6] The missing interview is: 20140524 Laumann

[7] The five development interviews are: 20140614 Maas, 20140802 Dressler, 20150124 Wendt, 20150404 Wagenknecht and 20151024 Peter. The five test interviews are: 20140517 Giegold, 20140927 Lemke, 20141011 Özoguz, 20150110 Bentele and 20150620 Münch. The rest of the documents make up the training data.

**DIRNDL** The DIRNDL anaphora corpus with updated bridging annotations was downloaded from the webpage[8]. We adopt the official train-development-test split.

### 5.2 Evaluation metrics

The evaluation of bridging resolution is computed using the widely known precision and recall measures (and the harmonic mean between them, F1). Additionally, we consider an antecedent correct if the predicted antecedent is one of the mentions in the coreference chain of the gold antecedent. For this, we take into account gold coreference chains. For optimisation, we use the development sets[9], and we report performance on the test set, if not indicated otherwise.

## 6 Baseline

In order to better judge how well the rule-based system performs, we create a baseline for anaphor and antecedent prediction. We first filter out all coreferent markables as annotated in the gold-standard. The baseline predicts a markable to be a bridging anaphor if it contains a definite article and is not modified by a prepositional phrase (PP), an adjective or does not contain a demonstrative pronoun (pre-processing is exactly the same as for the rule-based system, which we will describe later). The antecedent is then the subject of the previous sentence.

The baseline reflects the common ground that bridging anaphors are usually short, unmodified NPs and their antecedents usually appear in the previous sentence (cf. Hou, 2016b). The results of the baseline for DIRNDL and GRAIN are reported in Table 2 and 3.

The baseline achieves good performance for anaphor detection, suggesting that many bridging anaphors are indeed unmodified NPs, more so for GRAIN than for DIRNDL. The high recall is expected since the baseline suggests many candidates to be an anaphor, independent of other properties of the candidate. As a consequence, the precision is very low. The poor performance on the full prediction task is not surprising: Even though the antecedent often occurs in close proximity of its anaphor and subjects are the most preferred grammatical role, it is not necessarily the subject in the previous sentence.

| | Precision | Recall | F1 |
|---|---|---|---|
| Anaphor Rec. | 12.6% | 65.1% | 21.1% |
| Bridging Res. | 0.5% | 2.3% | 0.8% |

Table 2: Baseline results for anaphor detection and full bridging resolution on the test set of DIRNDL.

| | Precision | Recall | F1 |
|---|---|---|---|
| Anaphor Rec. | 15.8% | 69.8% | 25.9% |
| Bridging Res. | 0.4% | 1.6% | 0.6% |

Table 3: Baseline results for anaphor detection and full bridging resolution on the test set of GRAIN.

## 7 A rule-based approach

In this section, we describe our rule-based approach to bridging resolution. For this, we adapted the approach by Hou et al. (2014) to German. The system consists of three parts: (i) pre-processing, (ii) rule application and (iii) post-processing. For a more detailed explanation of the adaptation process, please refer to the supplementary material[10].

### 7.1 Pre-processing

We extract all gold markables of the information status annotation as our set of gold markables.

As potential bridging anaphor candidates, we filter out a number of noun types, as they are not considered bridging anaphors: all pronouns, indefinite expressions, proper names as well as markables which have embedded NPs and NPs whose head has appeared before in the document (as an approximation for coreferent anaphors). We also investigate the role of coreference information, as described in Section 7.3.1.

### 7.2 Rules

We implemented and adapted to German all eight rules as proposed by Hou et al. (2014). The input to the rules are the extracted markables. Each rule then proposes bridging pairs, independently of the other rules. The rules are summarised in Table 4. Some of the rules use the concept of semantic connectivity and argument-taking ratio, which we also adapted. The main idea behind the concept of semantic connectivity between two words can be ap-

[8] www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/dirndl.en.html

[9] From now on, we use the term *development set* for the combination of the development set and the training set of the respective corpus. By combining the two sets, we ensure a higher variety of bridging phenomena for tuning our system.

[10] www.ims.uni-stuttgart.de/institut/mitarbeiter/roesigia/bridging-resolution-german-supplementary.pdf

| Rule | Example | Anaphor | Antecedent search | Window |
|---|---|---|---|---|
| 1 | A white woman's house ← The basement | building part | semantic connectivity | 2 |
| 2 | She ← Husband David Miller | relative | closest person NP | 2 |
| 3 | The UK ← The prime minister | GPE job title | most frequent GEO entity | – |
| 4 | IBM ← Chairman Baker | professional role | most frequent ORG NP | 4 |
| 5 | The firms ← Seventeen percent | percentage expression | modifying expression | 2 |
| 6 | Several problems ← One | number/indefinite pronoun | closest plural, subject/object NP | 2 |
| 7 | Damaged buildings ← Residents | head of modification | modifying expression | – |
| 8 | A conference ← Participants | arg-taking noun, subj pos. | semantic connectivity | 2 |

Table 4: Overview of rules in Hou et al. (2014). For details, please refer to the supplementary material of this paper or the original paper.

proximated by the number of times two words occur in a N PREP N pattern. We computed the semantic connectivity scores using the SdeWaC corpus (Faaß and Eckart, 2013), a web corpus of 880 M tokens. The argument-taking ratio is a measure that describes the likelihood of a noun to take an argument. We derive the number of times in which a noun takes an argument automatically, by defining a number of patterns of modification (e.g. PP-postmodification, possessive modification), again using the SdeWac corpus. For a more detailed description, please refer to the original paper or the supplementary material of this paper.

#### 7.2.1 New rules

In addition to adapting the rules from the English system to German, we also added a couple of new rules, which are tailored to our domain of news and interviews.

**Rule 9: Country-related** It is common in our data that a country is introduced into the discourse and then a country-related entity is picked up as a bridging anaphor. Note that by country we mean both geographical locations as well as political entities.

(5) **Die Regierung** → Australien
(the government → Australia)

(6) **Die Westküste** → Japan
(the west coast → Japan)

We therefore introduce a new rule: If the anaphor is a non-demonstrative definite expression without adjectival or nominal pre-modification and without PP post-modification that occurs on our list of country parts, we search for the most salient country. Salience is determined by frequency in the document, with the exception of the subject in the very first sentence, which overrides frequency in terms of salience. The list of country parts consists of terms like *Regierung (government), Einwohner (residents)*, etc.

**Rule 10: High semantic connectivity** Rule 10 is similar to Rule 8 in Hou et al. (2014), but without the constraint that the anaphor has to be in subject position. However, it must be a non-modified NP or PP. If the semantic connectivity score to a previously introduced mention is higher than a certain threshold (15.0 in our experiments), it is proposed as the antecedent. The antecedent should appear in the last four sentences. The feature is designed to capture more general cases of bridging by looking for a high semantic connectivity between the anaphor and the antecedent.

**Rule 11: Political topics** This is a domain specific rule, based on the observation that many bridging anaphors in DIRNDL and GRAIN are related to political issues.

(7) **Parteivorsitzende** → die Grünen
(party leaders → the Green Party)

We obtain a list of nouns of the political domain from GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010). A markable is considered as an anaphor, if its head occurs in this list. Additionally, markables modified by adjectives or PPs are excluded. The antecedent is chosen by taking the markable with the highest semantic connectivity in the previous four sentences.

**Rule 12: Exclusion of `r-unused-known`** The evaluation of the baseline in Section 6 has shown that bridging anaphors are generally short and not modified by adjectives or PPs. Since we remove coreferent and indefinite expressions as possible anaphor candidates, the only other information status categories that frequently contain such expressions are `r-bridging` and `r-unused-known`. In Riester and Baumann (2017), the label `r-unused-known` is used for definite expressions which are generally known to the annotator. Rule 12 is identical to Rule 10, but aims to exclude such markables by only considering markables which only occur once in a document. The intuition is that known expressions are more salient and potentially occur multiple times in a discourse, while bridging anaphors are unique with respect to their context.

(8) im Internet ... im Internet ... im Internet
(on the Internet ... on the Internet ...
on the Internet)

(9) **Den Haken** → Das Kästchen
(the tick mark → the check box)

In the above examples, taken from one exemplary document, *Internet* appears three times in the whole document, while *Haken* only appears once. *Internet* is labeled as `r-unused-known`, since it is a generally known entity, while *Haken* is a bridging anaphor. Thus, in this case, Rule 12 will exclude all occurrences of *im Internet* as a potential bridging anaphor.

**Post-processing** In order to avoid conflicts of rules predicting different antecedents for the same anaphor, rule precision is evaluated on the development set. The rules are then ordered by precision and applied to the test set in descending order. Thus, a rule with a higher precision gets precedence over a rule with lower precision. The maximal sentence distance of the respective rules is also trained on the development set.

### 7.3 Results on DIRNDL

Table 5 shows the performance for both anaphor detection and full bridging resolution. As mentioned above, the performance was optimised on the development set and tested on the test set. Obviously, the scores for anaphor detection are higher, as the task of full bridging resolution predicts antecedents for the previously determined bridging anaphors. If all predicted antecedents are correct, the performance of full bridging resolution and anaphor detection are the same, which is of course not the case in our experiments.

| | Precision | Recall | F1 |
|---|---|---|---|
| Test set | | | |
| Anaphor Rec. | 26.0% | 18.9% | 21.9% |
| Bridging Res. | 16.3% | 11.6% | 13.6% |
| Dev set | | | |
| Anaphor Rec. | 47.6% | 19.0% | 27.2% |
| Bridging Res. | 26.7% | 10.5% | 15.1% |
| Whole set | | | |
| Anaphor Rec. | 39.1% | 19.1% | 25.6% |
| Bridging Res. | 22.2% | 10.7% | 14.4% |

Table 5: Performance of the rule-based system on DIRNDL.

| | Precision | Recall | F1 |
|---|---|---|---|
| Test set | | | |
| Anaphor Rec. | 45.5% | 15.9% | 23.5% |
| Bridging Res. | 22.7% | 7.9% | 11.8% |
| Dev set | | | |
| Anaphor Rec. | 29.4% | 15.2% | 20.0% |
| Bridging Res. | 17.4% | 9.0% | 11.9% |
| Whole set | | | |
| Anaphor Rec. | 32.1% | 15.3% | 20.7% |
| Bridging Res. | 18.3% | 8.8% | 11.9% |

Table 6: Performance of the rule-based system on GRAIN.

On the test set, the system achieves an F1 score of 21.9% for anaphor detection and 13.6% for bridging resolution. The precision is always higher than the recall, which is due to the focus on high precision rules. We also tested how the system performs on the development set, also displayed in Table 5. Overall, the performance is higher, which was to be expected, since the system was optimised on this subset. However, the differences are not very large, suggesting that the system is not overfitting to the development set and the rule ordering and maximum sentence distances that it learned also work well on unseen data. Table 5 also presents the performance for the whole data set, for both anaphor detection and full bridging resolution[11].

Most of the rules transferred from the English bridging resolver do not predict any bridging pairs in our data. For some cases, this can be explained by the different bridging definitions (e.g. no indefinite bridging anaphors in our data). Rule 6, for example, which is designed to resolve anaphors containing a number expression or indefinite pro-

[11] These values are later used as references when we investigate possible sources of error for our system.

nouns, cannot propose any correct pairs due to guideline differences.

Of course, ISNotes (Markert et al., 2012), the corpus on which the experiments in the English bridging resolver were based on, and DIRNDL are also of slightly different domains (news text in ISNotes vs. radio news in DIRNDL), which might explain some of the differences.

Table 7 shows the performance of the single rules when being applied to DIRNDL. From the original English system, only Rule 4 (GPE job titles) and the very general Rule 8 (which is based on semantic connectivity) fire. Our new rules also predict pairs: While Rule 9 (country-related) is rather specific and has a high precision, Rule 10 proposes a lot of pairs, thus increasing the recall. Rule 12 is highly similar to Rule 10, and, for DIRNDL, does not seem to help more than Rule 10, indicating that filtering out `r-unused-known` entities was not successful for DIRNDL. Rule 11 (political topics) is very specific and similarly to Rule 9, it is also based on lexical lists of potential bridging anaphors, but cannot achieve a similarly high precision.

#### 7.3.1 Bridging resolution with gold coreference

To test the effect of coreference information, we also run the system without filtering out coreferent anaphors. In Table 9, we show that, as expected, the precision and, as a result, the F1 score are significantly higher in the setting with coreference[12].

### 7.4 Results on GRAIN

In order to test the generalisability of the findings, we also report results on GRAIN. The results of the system's performance on GRAIN are shown in Table 6. For anaphor detection, the system performs better on GRAIN than on DIRNDL with an F1 score of 23.5%, compared to 21.9% for DIRNDL. However, this effect was only observed on the test data, not on the development set. Overall, the performance on GRAIN for full bridging resolution is notably and consistently lower than on DIRNDL (11.8% vs. 13.6%). The data sets for GRAIN also seem to be fairly distributed in terms of bridging anaphors, since all F1 values are rather close together.

While 97.9% of all nouns appearing in DIRNDL have an argument-taking ratio score and 45.9% of the noun-noun combinations have a semantic connectivity score, we find that in GRAIN, 98.3% of all nouns have an argument-taking ratio score, but only 24.0% of the noun-noun combinations have a semantic connectivity score. We believe that this is one of the reasons for the overall lower score on full bridging resolution. Another reason could be that while the radio news in DIRNDL are scripted and have prototypical topics such as politics, the weather, etc., GRAIN contains spontaneous speech of very diverse topics.

Results on the precision of the single rules are displayed in Table 8. Overall, the rules perform worse than for DIRNDL. In addition to that, Rule 11 does not seem to work for GRAIN very well.

## 8 Oracle lists

For GRAIN, finding the correct antecedent for a bridging anaphor is noticeably more difficult than for DIRNDL. In order to investigate why this is the case, we do some experiments using oracle lists to find antecedents in both GRAIN and DIRNDL. An oracle list represents a ranked suggestion of antecedents for an anaphor, with the most likely antecedent on top. Despite the fact that the rules in our system only predict one antecedent, we can change them so that they predict several antecedents. For this, we use the antecedent search technique of the respective rule and extend it to predict several candidates, instead of just one antecedent. For example, some of the rules are based on distance (often combined with a restriction, e.g. the closest organisation). Instead of predicting only the closest organisation, we can now come up with a list of organisations, ranked by distance. Other rules are based on the semantic connectivity scores, where we can then use the scores to create the list of potential antecedents. Note that we do not change the rules, nor do we involve any sort of re-ranking: we simply use the rule's search technique to create a list of antecedents, rather than a single antecedent[13]. This way, we can evaluate

[12] We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.05 level.

[13] To avoid ties, we perform simple modifications in order to influence the ranking. For example, Rule 3 also ranks according to document frequency of candidates, but we take into account the sentence and word distance, to penalise candidates which are further away from the anaphor. In case a rule already predicted a candidate to be a potential antecedent for a previous anaphor, we push these candidates higher on the ranking by adding a fixed value. This is meant to take into account the fact that antecedents are often the antecedent of multiple anaphors (cf. Hou (2016b)'s findings on sibiling anaphors).

| | Anaphor detection | | | Full bridging resolution | | |
|---|---|---|---|---|---|---|
| Rule | Correct | Wrong | Precision | Correct | Wrong | Precision |
| Rule 4: | 4 | 0 | 100.00% | 0 | 4 | 0.0% |
| Rule 8: | 23 | 24 | 48.9% | 8 | 39 | 17.0% |
| Rule 9: | 27 | 7 | 79.4% | 22 | 12 | 64.7% |
| Rule 10: | 50 | 63 | 44.2% | 20 | 93 | 17.7% |
| Rule 11: | 10 | 10 | 50.0% | 4 | 16 | 20.0% |
| Rule 12: | 50 | 63 | 44.2% | 20 | 93 | 17.7% |

Table 7: Rule precision on the development set of DIRNDL.

| | Anaphor detection | | | Full bridging resolution | | |
|---|---|---|---|---|---|---|
| Rule | Correct | Wrong | Precision | Correct | Wrong | Precision |
| Rule 1: | 1 | 5 | 16.6% | 1 | 5 | 16.6% |
| Rule 4: | 0 | 2 | 0.0% | 0 | 2 | 0.0% |
| Rule 8: | 10 | 16 | 38.5% | 3 | 23 | 11.5% |
| Rule 9: | 15 | 17 | 46.9% | 13 | 19 | 40.6% |
| Rule 10: | 7 | 30 | 18.9% | 3 | 34 | 8.1% |
| Rule 11: | 1 | 13 | 7.1% | 0 | 14 | 0.0% |
| Rule 12: | 6 | 28 | 17.6% | 3 | 31 | 8.8% |

Table 8: Rule precision on the development set of GRAIN.

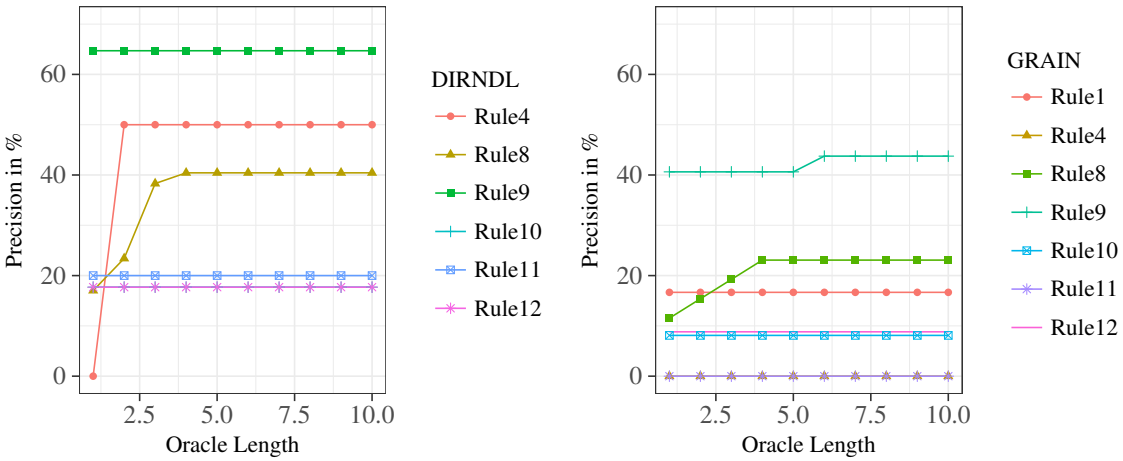


Figure 1: Performance of rules on the development set for DIRNDL and GRAIN, using different lengths of oracle lists.

| Setting | Precision | Recall | F1 |
|---|---|---|---|
| No coref | 10.7% | 11.6% | 11.1% |
| Gold coref | 14.9% | 11.6% | 14.4% |

Table 9: Bridging resolution with different types of coreference information in DIRNDL (Gold markables).

the antecedent search strategies of the respective rules.

Figure 1 shows the precision for each rule based on the length of the oracle list, evaluated on the development set. We can see that the rules benefit from the oracle lists to a different extent. Rule 9 in DIRNDL is not changing its precision, suggesting that its performance is already quite good and all correct antecedents are already ranked on top of the oracle. Other rules like Rule 4 or 8 benefit a lot, indicating that the correct antecedents are generally in the scope of the rule, but simply not ranked high enough. Rule 4 and 11 in GRAIN stay at 0% precision. This means that these rules are not able to capture the correct antecedents at all.

In Figure 2, the overall performance of the system on the whole dataset is shown, dependent on the oracle length. Both datasets benefit from the oracle lists, but especially GRAIN could benefit from re-ranking the oracle lists in order to push the correct antecedent higher. Overall improvement through re-ranking is however limited, since many rules are restricted in their search for an antecedent by the maximum sentence distance. The fact that some of the rules cannot show their full

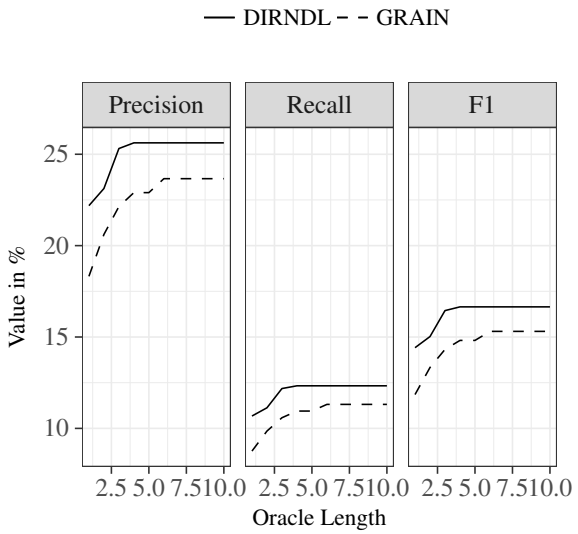


Figure 2: Performance of the rule-based system on the whole data set for DIRNDL and GRAIN, using different lengths of oracle lists.

potential even for a higher oracle list length suggests that these rules have no access to the correct antecedent at all and need to be revised.

## 9 Variable Importance

We investigated different machine learning techniques, but due to the small amount of data, the results were lower than for the rule-based approach and thus not shown here. However, we use machine learning in order to evaluate the importance of different features that were also used in the rule-based system. Doing so, we get a better understanding of what features are actually beneficial for the rule-based system.

We look closer at the prediction power of a few selected features. These are the length and number of words of the anaphor, the POS of the head of the anaphor, the anaphor's argument taking ratio, the sentence distance from the anaphor, the POS and named entity (NE) category of the head of the antecedent, its length and word count and the semantic connectivity.

We report variable importance values using the random forest technique (Ho, 1995) with 10-fold cross validation on GRAIN. Variable importance is estimated by leaving out a single feature for prediction and evaluating the decrease in performance for the random forest classifier. Table 10 shows the results for anaphor detection and bridging resolution.

It becomes clear that semantic connectivity, the argument-taking ratio of the anaphor and the length in characters of the anaphor/antecedent are overall good predictors. This substantiates the use of these features, since the rule-based system makes extensive use of them. However, coverage and computation of semantic connectivity should be improved in order to obtain better results of antecedent detection for GRAIN.

| Feature | Variable Importance |
|---|---|
| SemanticConnectivity | 32.2 |
| AnaCharLength | 31.6 |
| AnteCharLength | 30.5 |
| AnaArgTakingRatio | 29.3 |
| AnteWordCount | 25.9 |
| AnaWordCount | 22.5 |
| SentDist | 14.9 |
| AnteHeadPOS | 5.9 |
| AnteHeadNE | 5.8 |
| AnaHeadPOS | 3.3 |

Table 10: Variable importance estimated with a random forest classifier on GRAIN.

## 10 Conclusion

We have presented an analysis of bridging in two available corpora for German, DIRNDL and GRAIN. We have implemented a baseline for bridging resolution, which achieved good results for anaphor detection, indicating that short, unmodified NPs are good bridging anaphor candidates, but resulting in poor performance for bridging resolution. We have also presented a rule-based system following Hou et al. (2014), which has achieved reasonable results on both corpora. Oracle lists have shown the potential of the single rules if they were better at finding the correct antecedent, which could be exploited in a re-ranking approach. The features and information used by the rule-based system seem to be promising, but could still be improved and extended.

## Acknowledgments

We would like to thank Arndt Riester for his valuable comments as well as the anonymous reviewers for their insightful remarks. This work was funded by the Collaborative Research Center SFB 732, Project A6.

## References


Stefan Baumann and Arndt Riester. 2012. Referential and Lexical Givenness: semantic, prosodic and cognitive aspects. In Gorka Elordieta and Pilar Prieto, editors, *Prosody and Meaning*, number 25 in Interface Explorations. Mouton de Gruyter, Berlin.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC 2014, pages 3222–3228.

Aoife Cahill and Arndt Riester. 2012. Automatically acquiring fine-grained information status distinctions in German. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 232–236, Seoul.

Herbert H. Clark. 1975. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A discourse information radio news database for linguistic analysis. In *Linked Data in Linguistics*, pages 65–76. Springer.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC - A corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web - 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25-27, 2013. Proceedings*, pages 61–68.

Yulia Grishina. 2016. Experiments on bridging across languages and genres. In *Proceedings of the first Workshop on Coreference Resolution Beyond OntoNotes (NAACL-HLT)*, pages 7–15, San Diego, USA.

Udo Hahn, Michael Strube, and Katja Markert. 1996. Bridging textual ellipses. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 496–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.

John A Hawkins. 1978. *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Crook Helm.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT – The GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, LREC 2010, pages 2228–2235.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pages 278–282, Montreal, QC.

Yufang Hou. 2016a. Incremental fine-grained information status classification using attention-based LSTMs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1880–1890, Osaka, Japan.

Yufang Hou. 2016b. *Unrestricted Bridging Resolution*. Ph.D. thesis, Heidelberg University.

Yufang Hou, Katja Markert, and Michael Strube. 2013a. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 814–820, Seattle, USA.

Yufang Hou, Katja Markert, and Michael Strube. 2013b. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, USA.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2082–2093, Seattle, USA.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 795–804. Association for Computational Linguistics.

Katja Markert, Michael Strube, and Udo Hahn. 1996. Inferential realization constraints on functional anaphora in the centering model. In *In Proc. of the 18 th Annual Conference of the Cognitive Science Society; La*, pages 609–614.

Altaf Rahman and Vincent Ng. 2012. Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 798–807, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2010. A typology of near-identity relations for coreference (nident). In *LREC*.

Arndt Riester and Stefan Baumann. 2017. The RefLex Scheme – Annotation guidelines. SinSpeC. Working papers of the SFB 732 Vol. 14, University of Stuttgart.

Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Faleńska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn. 2018. German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC 2018.

Sidney Siegel and N. John Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. McGraw-Hill, Berkeley, CA.