1 Supplementary material: 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.

1.1 Pre-processing

We extract all markables with 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:

- Pronouns: all pronouns are excluded as they are typically either pleonastic or coreferent with an already introduced entity.
- Indefinite expressions: all indefinite markables should, as stated in the guidelines, not be bridging anaphor candidates. We use a set of definite determiners to determine the definiteness of the markables.
- Proper names: proper names are also definite, but are not suited as bridging anaphors as they typically occur as expressions of the category `unused/mediated`. NPs containing embedded proper names can of course be of the category bridging and should not be excluded.
- Markables whose head has appeared before in the document: this is meant as an approximation for coreferent anaphors.
- NPs that have embedded NPs. In practice, this leads to the exclusion of long NPs that have embedded markables, e.g.

  (1) unter dem Deckmantel der zivilen Nutzung der Nuklearenergie
  (under the guise of civilian use of nuclear energy)

  These expressions are typically of the information status category `unused-unknown`.

Filtering of bridging antecedent candidates When using predicted markables, overlapping markables can be extracted. To overcome this, we filter out embedded NEs that occur in NPs or PPs from the set of potential antecedents , but only if the NP or PP differs from the NE only in the form of a determiner, preposition or a pre-modifying noun.

(2) Der Iran

(3) Im Iran

(4) Bundesaußenminister Steinmeier

Not excluded are embedded NPs in other constructions, e.g. in

(5) auf Wunsch Spaniens

1.2 Rules

We have 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 1[1]. Some of the rules use the concept of semantic connectivity and argument-taking-ratio, which we describe here in more detail, because the computation differs from the one in the original paper.

1.2.1 Semantic connectivity

The semantic connectivity goes back to the NP of NP pattern in Poesio et al. (2004) and was extended to a more general preposition pattern in Hou et al. (2014). The main idea is that semantic connectivity between two words can be approximated by the number of times two words occur in a N PREP N pattern (or in our extended version: in a noun preposition (optional: determiner and adjective) noun pattern)). This means that two nouns like *Sand* and *Strand* (sand and beach) have a high semantic connectivity because they often occur as *Sand am Strand* (sand on the beach), whereas other nouns do not appear often in such a construction and are therefore not highly semantically connected.

Following Hou et al. (2014), we compute the Dunning root log-likelihood ratio (Dunning, 1993) as a measure of strength of association. In contrast to Hou et al. (2014), we do not limit prepositional patterns to the three most common prepositions for a noun, but count every N PREP N pattern. We take the SdeWaC corpus (Faaß and Eckart, 2013),

[1] For a more detailed description, please refer to the original paper.

Example	Anaphor	Antecedent search
Rule1: The basement→ a white woman's house	building part	semantic connectivity, window 2
Rule2: Husband David Miller → she	relative	closest person NP, window 2
Rule3: The prime minister → the UK	GPE job title	most frequent GEO entity
Rule4: Chairman Baker → IBM	professional role	most frequent ORG NP
Rule5: Seventeen percent → the firms	percentage expression	modifying expression, window 2
Rule6: One → several problems	number or indefinite pronoun	closest plural, subject/object NP
Rule7: Residents → damaged buildings	head of modification	modifying expression, window 2
Rule8: Participants → a conference	arg-taking noun, subj position	semantic connectivity

Table 1: Overview of the rules in Hou et al. (2014).

a web corpus of 880 M tokens, to compute the semantic connectivity for all combinations of nouns that occur in this prepositional pattern in the corpus. This way, we not only compute the numbers for nouns in DIRNDL or GRAIN, but also for other nouns, making the approach applicable for new texts.

In contrast to English, German has many one-word compounds, like *Hüpfkind (jumping kid), Schreikind (screaming kid)*. Many of these are infrequent, thus leading to sparsity issues. To overcome this, we apply the compound splitter Compost (Cap, 2014), and compute the semantic connectivity for the heads of the respective compounds. This reduces the number of pairs from 12,663,686 to 8,294,725.

1.2.2 Argument-taking ratio

The argument-taking ratio is a measure that describes the likelihood of a noun to take an argument. The idea behind this concept is that there are words which are often used generically, like *children*. Others, like *husband* mostly occur with a modifier (e.g. *husband of*), and are thus more likely candidates for bridging. In the English bridging resolver, this was computed with the help of the NomBank annotations. These manual annotations list, for every occurrence in the WSJ corpus, the arguments of the nouns. To compute the argument-taking ratio, one then simply has to divide the number of NomBank annotations for one noun by the total frequency of the noun in the corpus. This is only possible because both the ISNotes and the NomBank annotation were performed on the same corpus. For other corpora or texts, we need to derive the number of cases in which the noun takes an argument automatically.

To do this, we define these patterns of modification:

1. PP-postmodification :
   N_{target} PREP (Det) (ADJ)* N
   Türen im Haus (doors in the house)
2. NPgen-postmodification:
   N_{target} (Det) (ADJ)* N
   die Kinder der Frau (the woman's kids)
3. Possessive pre-modification:
   POSS N_{target}
   Ihr Ehemann (her husband)

We then divide the frequency of a noun in these constructions by the total frequencies of the noun in a large corpus. Again, we use the SdeWaC corpus to derive the argument-taking ratio scores. As in the computation of the semantic connectivity scores, we run into sparsity issues due to infrequent compounds. Thus, we also apply the compound splitter, to get more stable ratios. The argument-taking ratios are compiled for the head of the noun, if a compound split exists. This reduces the number of nouns from 5,527,197 to 2,335,293.

Rule 1: Building-part-of The anaphor is a part of a building (e.g. window, room, etc.) and is not pre-modified by a common or proper noun. The antecedent is selected as the one with the highest semantic connectivity in the same or the previous two sentences.

(6) **Die Fenster** – <u>im Zimmer</u>

There is no noun in the DIRNDL corpus that is present on the building list, i.e. this rule is not particularly suited for our domain. It is left in anyway as it could be relevant in other data.

Rule 2: Relative person NPs The anaphor is on a list of relative nouns (e.g. child, son, husband, etc.), its argument taking ratio is greater than 0.4 (meaning that it is not used generically, i.e. in *children like toys*, but typically appears with an argument *husband of XY*. It is not modified by an adjective, a noun or a PP.

Antecedents must be in the same sentence or the two previous ones and either a proper noun and not

a location, or a named entity tagged as person, or a personal pronoun except second person *du*.

(7) **Ihr Mann** – Martha

Rule 3: GPE job titles The anaphor is on a list of official job titles for a country (e.g. commissioner, secretary, etc.). It is not modified by a country modification *der argentinische Außenminister* and not modified by a PP or an organisation.

The antecedent is the most salient geopolitical entity in the document. Salience is determined by frequency in the document. In case of ties, the closest is chosen.

(8) **Der Außenminister** – Deutschland

Rule 4: professional roles

(9) **CEO Peter Müller** – IBM

(10) **Der Vorstand** – der SPD

The head of the anaphor appears on a list of professional roles, like *Manager, Arzt* and is not modified by a country, a PP, a proper name or an organisation. The most salient antecedent is chosen within the last four sentences. Salience is determined by frequency in the document.

Rule 5: Percentage expressions

(11) **5** % – der Deutschen.

The anaphor is a percentage expression containing % or "Prozent". As antecedent, the modifier expression of another percentage expression is chosen, e.g. *der Deutschen* in *10* % *der Deutschen*. This rule is not applicable to DIRNDL/GRAIN as these percentage expressions are indefinite.

Rule 6: Other set members This rule is not applicable for our data as it is designed for indefinite anaphora. It is left unimplemented in the resolver, in case one wants to implement it for other corpora.

Rule 7: argument-taking ratio 1 The anaphor is a common noun phrase (non-modified) with an argument-taking ratio over 0.4.

The antecedent is determined by finding the closest similar modification in the document. For details, refer to the original paper.

Rule 8: argument-taking ratio 2 The anaphor is a definite, non-modified expression in subject position (where it is likely to either be coreferent or bridging) with an argument-taking ratio over 0.4.

The antecedent is chosen as the entity with the highest semantic connectivity in the last three sentences.

1.2.3 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 part-of It is common in our data that a country is introduced into the discourse and then a part of the country is picked up later as a bridging anaphor.

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

(13) **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, which can be found 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.

(14) **Halbzeit** → Die große Koalition
(halftime → the grand coalition)

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 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.

Post-processing The rules are ordered and applied according to their precision.

References

Fabienne Cap. 2014. Morphological processing of compounds for statistical machine translation. Dissertation, Institute for Natural Language Processing (IMS), University of Stuttgart.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

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.

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.

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.

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 *EMNLP*, pages 2082–2093.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 143. Association for Computational Linguistics.

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