Ambiguity in German Connectives: A Corpus Study

Angela Schneider
Master Student
Department of Computational Linguistics
Universität Heidelberg / Germany
schneida@cl.uni-heidelberg.de

Manfred Stede
Applied Computational Linguistics
EB Cognitive Science
Universität Potsdam / Germany
stede@uni-potsdam.de

Abstract

For deriving information on text structure (in the sense of coherence relations holding between neighbouring text spans), connectives are the most useful source of evidence on the text surface. However, many potential connectives also have a non-connective reading, and thus a disambiguation step is necessary. This problem has received only relatively little attention so far. We present the results of a corpus study on German connectives, designed to estimate the magnitude of the ambiguity problem and to prepare the development of disambiguation procedures.

1 Introduction

Connectives are a central source of information for *discourse parsing*, i.e., the identification of structural relationships between sentences or clauses in text. As with many lexical items, however, there is an ambiguity problem that needs to be resolved before such structural information can be exploited. We have conducted a corpus study in order to determine the magnitude of the ambiguity problem for German connectives, and to pave the way for implementing effective disambiguation procedures. The results will be presented in Section 3. Before, we briefly introduce the notions of discourse parsing and connectives in the remainder of this section, and discuss related work in Section 2.

1.1 Discourse parsing

Coherence relations are often used to model the coherence of texts, and sometimes to also ascribe structure to it. For instance, many researchers would analyze the discourse *The hotel is nice and clean. But I think it is much too expensive* as an instance of the *Concession* relation holding between the two sentences. Certain theories of discourse structure then take the observation of such relations a step further and postulate that, by means of recursive application of such relations, a structural description can be produced as a model of the text's internal coherence (e.g., (Polanyi, 1988) (Mann and Thompson, 1988), (Asher and Lascarides, 2003)). The step of automatically building these descriptions is commonly called *discourse parsing* (see, e.g., (Marcu, 2000)).

Regardless whether the aim is to derive a full text structure or to merely identify local structural configurations at selected points of a text, the identification of coherence relations is made much easier when explicit connectives are present: words that – more or less explicitly – signal the presence of a coherence relation.

1.2 Connectives

Connectives are non-inflectable, closed-class lexical items that denote two-place relations, i.e., they need two arguments in order to be used felicitously (Pasch et al., 2003). Traditional grammars often group connectives together according to their semantic function (e.g., contrastive, temporal, causal); the mapping to a coherence relation can be seen as a discourse-level extension of that grammatical analysis. Syntactically, connectives are not homogeneous; we find coordinating conjunctions (e.g., *and, but*), subordinating conjunctions (e.g., *while, because*), and discourse ad-

verbials (e.g., *therefore, still*). Some researchers also include certain propositions such as *due to* or *despite*.

The mapping from connective to coherence relation is non-trivial for three different reasons. (i) Connectives vary in their specificity, ranging from very clear ones (*although*) to vague ones (*and*). (ii) Some connectives are ambiguous between different semantic readings (or, coherence relations), such as *while*, which can be temporal or contrastive. (iii) Some words have a connective and a non-connective reading, such as *since*, which can be a subordinating conjunction or a preposition without discourse function (*She has been a widow since 1983*).

In the following, we address only problem (iii), and in particular study it for the German language. Our method, however, should be applicable to other languages just as well.

2 Related work

There are quite a few papers dealing with connective ambiguity of the kind described in point (ii) above, but regarding the ambiguity between a connective and non-connective reading for English words, we are aware of only (Pitler and Nenkova, 2009). In this study, Pitler and Nenkova investigate to what extent syntactic information is useful in solving this ambiguity problem.

For German, (Bayerl, 2004) had presented a pilot study on disambiguating the potential connective *wenn*, also using various syntactic features, but her focus was on ambiguity (ii). Regarding (iii), (Dipper and Stede, 2006) suggested the approach to incrementally retrain a POS tagger with non-/connective features, but their experiment was restricted to nine words. To our knowledge, there is no study of the "bigger picture" of connective ambiguity yet.

3 Corpus study

The first problem is to identify the set of ambiguous words: those that have a connective and a non-connective reading. For German, such a set has been determined in earlier work by (Dipper and Stede, 2006) who gave a list of 41 German connectives that can be ambiguous; these are listed below in Table 1.

For each of these words we now collected a small corpus consisting of 200-250 sentences taken from the *DWDS-Kernkorpus*[1] Then we manually annotated all sentences as to the candidate words having a connective or a non-connective reading.[2] In this way, we ended up with 1-200 sentences for each reading of each potential connective. Since the sentences were collected randomly, the distribution of non-/connective readings can be interpreted as approximating the actual distribution in language. Table 1 includes this information. Notice that some words have a very skewed distribution so that a "default" reading could be assumed (*auch, nur, wie* and others), whereas most show a balance of the two readings, so that disambiguation is indeed non-trivial. In the table, we also give the raw frequency of the words' occurrences in the DWDS corpus (which is the parameter for sorting the table). For the highly infrequent connective readings of *auch* and *nur* we subsequently searched for another 50 occurrences, in order to have enough material for the disambiguation steps described below.

3.1 Standard POS tagging

An obvious first idea is to explore standard POS tagging for the disambiguation problem (iii) as stated in section 1.2. Therefore, we tagged our data with the TreeTagger (Schmid, 1994), which uses the STTS Tagset[3]. However, it turns out that POS tags disambiguate only eight potential connectives with an f-score > 0.75 (which we take as a threshold for "acceptable" performance). Tables 2 and 3 show the respective tags and their precision and recall.

A side result of our analysis is that for some POS tags, general rules can be established; in partiuclar, words tagged with the tags PTKVZ, NN or NE are always non-connectives. Overall, however, the intermediate conclusion is negative: Standard POS tagging yields acceptable results only for eight of the 41 candidates. And, as Table 1 reveals, the eight "easy" words range in the

[1] http://www.dwds.de

[2] The criteria that were applied in making the annotation decisions are documented in (Schneider, 2012).

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.htm

middle of the frequency distribution, so that the problem at large can be solved only to a small extent by tagging.

3.2 POS-context disambiguation rules

For the remaining 33 potential connectives (where the POS-tag alone does not give enough information to disambiguate), our next step was to look at their immediate context and determine wheter POS tag patterns can be identified for making the decision. We found that for ten potential connectives, it is indeed possible to formulate a set of context-rules that use the POS-tags of the tokens directly before and after the potential connective, and which perform with an f-score > 0.75. These connectives and the context rules are shown in table 4. The underlined POS-tag is the one the potential connective is annotated with, while the POS-tags before and/or after describe the directly adjacent words. To merge some context rules together the following symbols are used: $ stands for any punctuation mark, VV.* for any full verb and V.FIN for any finite verb.

4 Summary and outlook

We provided a comprehensive analysis of the connective ambiguity problem in German and showed for our base set of 41 ambiguous words that

- a set of words has a relatively clear tendency to occur in either the connective or non-connective reading,
- for eight words, plain POS tagging yields good results for disambiguation, and
- for another ten words, fairly reliable POS-context patterns can serve to disambiguate the reading.

These results can serve as the basis for an implementation. For the remaining ambiguous words, a straightforward first step is to simply assume the majority reading as taken from Table 1. Alternatively, one could search more intensely for POS patterns by working with larger corpora and machine learning techniques. And another option to explore, of course, is to test whether syntactic chunking or even full parsing can be trusted to resolve the ambiguities. For English, this has been done by (Hutchinson, 2004), using the Charniak parser. In an interesting experiment, (Versley, 2010) suggested to project connective annotations from English onto German data, thus cirvumventing the problem of lacking (German) training data for machine learning. Working directly on the output of a German parser, however, has to our knowledge not been attempted yet.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Petra Bayerl. 2004. Disambiguierung deutschsprachiger Diskursmarker: Eine Pilot-Studie. *Linguistik Online*, 18.

Stefanie Dipper and Manfred Stede. 2006. Disambiguating potential connectives. In Miriam Butt, editor, *Proc. of KONVENS '06*, pages 167–173, Konstanz.

Ben Hutchinson. 2004. Mining the web for discourse markers. In *Proc. of LREC-04*, Lisbon.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press, Cambridge/MA.

Renate Pasch, Ursula Brauße, Eva Breindl, and Ulrich Herrmann Waßner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proc. of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc. of International Conference on New Methods in Language Processing*, pages 44–49, Manchester.

Angela Schneider. 2012. Disambiguierung von Diskurskonnektoren im Deutschen. Bsc. thesis, Dept. Linguistics, Universität Potsdam, March.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proc. of the Workshop on the Annotation and Exploitation of Parallel Corpora (AEPC)*, Tartu/Estland.

word	raw freq.	non conn. (of 200)	conn. (of 200)
und	1.550.931	121	79
als	382.237	183	17
auch	351.946	199	1
wie	268.878	188	12
so	263.201	163	37
nur	226.392	199	1
aber	219.248	55	145
dann	106.610	56	144
doch	92.350	120	80
da	88.776	110	90
denn	60.647	84	116
also	54.251	142	58
seit	39.670	179	21
während	39.320	98	102
darauf	32.630	187	13
dabei	27.616	181	19
allein	24.781	183	17
wegen	20.994	9	191
dafür	20.857	178	22
daher	18.874	16	184
sonst	18.139	144	56
statt	15.749	163	37
zugleich	15.093	119	81
allerdings	14.481	33	167
dagegen	13.388	52	148
ferner	12.622	18	182
trotz	10.921	20	180
darum	10.090	120	80
außer	10.084	185	15
soweit	8.884	31	169
entgegen	6.904	142	58
danach	6.351	85	115
wonach	3.042	186	14
worauf	2.840	103	97
weshalb	2.638	124	76
seitdem	2.396	139	61
womit	2.048	109	91
aufgrund	1.806	200	0
allenfalls	1.162	177	23
wogegen	614	54	146
nebenher	286	93	107
weswegen	188	111	89

Table 1: Ambiguous connectives with their raw frequencies and non-/connective distributions for 200 random occurrences

	POS = connective	Prec. in %	Recall in %
denn	KON	85.6	94.2
doch	KON	86.3	83.1
entgegen	APPO, APPR	97.9	65.7
seit	KOUS	77.8	84.0
seitdem	KOUS	81.0	77.0
trotz	APPR	99.5	100
während	KOUS	81.8	96.1
wegen	APPO, APPR	98.3	100

Table 2: POS tagging results for connective readings

	POS = non-connective	Prec. in %	Recall in %
denn	ADV	91.9	79.1
doch	ADV	90.2	92.1
entgegen	PTKVZ	85.7	99.3
seit	APPR	98.0	96.6
seitdem	PAV	90.2	92.1
trotz	NN	100	95.0
während	APPR	95.3	78.6
wegen	NN	100	60.0

Table 3: POS tagging results for non-connective reading

	Context-rules connective	Prec. in %	Recall in %	Context-rules non-connective	Prec. in %	Recall in %
also	$, ADV V.FIN $. ADV V.FIN V.FIN ADV	83.3	87.3	all else	95.0	93.2
auch	ADV VVFIN	98.0	78.1	all else	95.3	99.7
außer	$ APPR $, $ APPR KOUS	100	86.7	all else	98.9	100
da	$, ADV KON ADV KOUS	77.5	86.9	ADV PTKVZ	87.7	78.8
darum	all else	70.7	81.7	PAV $ PAV VV	82.5	89.7
nebenher	$. ADV VVFIN all else	88.2	83.3	ADV $ ADV VV.* ADV KON	81.8	87.1
nur	$. ADV VVFIN	100	74.2	all else	93.6	100
so	$, ADV KOUS $, ADV V.FIN KON ADV V.FIN	77.8	77.8	all else	95.1	95.1
sonst	$ ADV V.FIN	87.2	70.7	all else	89.6	96.1
soweit	$ ADV KOUS	98.9	92.3	all else	65.9	93.5

Table 4: Evaluation results for POS patterns