Disambiguating potential connectives

Stefanie Dipper and **Manfred Stede**
Angewandte Computerlinguistik
Inst. für Linguistik
Universität Potsdam
Karl-Liebknecht-Str. 24-25
D-14476 Potsdam-Golm
dipper|stede@ling.uni-potsdam.de

Abstract

Many discourse connectives also have non-discourse, or *sentential* readings. Therefore, for automatic discourse structure analysis, there arises a disambiguation problem even before the question of signalled discourse relation beomes relevant. We focus here on a set of nine German connectives and characterize the task of determining their discourse/sentential reading. Starting from an analysis of the utility of state-of-the-art PoS taggers, we describe a series of experiments with training the Brill tagger for identifying connectives. Our results indicate that there is a relatively simple baseline approach, which retraining the tagger can in turn improve on, but not very much.[1]

1 Introduction

Discourse connectives are closed-class lexical items that indicate the type of relationship between portions of text. As such, they figure prominently among the various *cohesive* devices in language. Connectives differ in terms of their specificity, though. For example, *however* signals a relation of contrast or concession, whereas *and* signals a very unspecific addition — which is nonetheless distinct from the mere juxtaposition of clauses without connective (see, e.g., Blakemore and Carston (2005)). As is well known, connectives do not form a syntactically homogeneous class, and moreover there is a fuzzy border to various kinds of phrasal expressions such as *in other words* or *to be more specific*. We discuss the problem of defining connectives in some more detail in Section 2, and also provide some figures to illustrate the range of the problem.

In computational linguistics, connectives have been employed by a variety of text generators, and more recently also in approaches to automatic text understanding, sometimes labelled 'rhetorical parsing'. Using shallow, surface-based methods, researchers have tried to build underspecified discourse structures from basically unrestricted text. These, in turn, have applications in tasks like question answering or text summarization. — This work is briefly described in Section 3, which also motivates the setup of our own approach to text understanding.

While the utility of analyzing connectives for such computational goals is undisputed, earlier research has to our knowledge largely neglected the problem of connective *disambiguation*. This means on the one hand ambiguity with respect to the signalled coherence relation, and on the other hand ambiguity as to the discourse function in general. It is the latter aspect that we focus on in this paper. Many words that can function as connectives also have a non-discourse, or *sentential* reading (in the terminology of Hirschberg and Litman (1994)). The English *but*, for instance, can signal a contrast or concession relation, or it can mean 'except', as in *Everybody but Peter attended the party.* While this is relatively unproblematic for text generation, it has important ramifications for text understanding: we do not want to hypothesize a coherence relation from a word that is in fact not used as a connective.

As a step of pre-processing before discourse structure analysis, we thus propose *connective tagging*: The decision whether an ambiguous word in some specific instance has a discourse reading or a sentential reading. We have used the Brill tagger (Brill, 1992) to learn a language model for German and handle connectives by rewriting the "standard" part-of-speech tags of potential discourse connectives to a 'DC' tag, which can then be utilized by subsequent analysis steps that try to build a discourse structure. Section 4 reports on our various experiments with training the Brill tagger to per-

[1]The research reported in this paper was financed by Bundesministerium für Bildung und Forschung, grant 03WKH22.

form this disambiguation; we compare a number of different training scenarios. The results show that automatic identification of non-connective readings yields good results, while identifying the connective reading is more difficult with our method.

2 Connectives and their ambiguity

The most comprehensive source of information on German connectives, Pasch et al. (2003), lists about 350 different entries and gives detailled syntactic characterizations. More generally, the authors propose five criteria for defining the notion of *connective*. A candidate word x has to fulfil all of these:

(M1) x cannot be inflected.

(M2) x does not assign case features to its syntactic environment.

(M3) The meaning of x is a two-place relation.

(M4) The entities related by the meaning of x are states of affairs ('Sachverhalte').

(M5) The entities related by the meaning of x can be expressed as finite clauses.

A different approach to defining the notion was taken by Knott (1996), who proposed a procedural test for identifying connectives in English sentences. This test, however, is not straightforwardly applicable to German, as argued by Grote (2003), who suggested an extension of Knott's procedure (p. 85). For our present purposes, however, the characterization by (M1)–(M5) is sufficient, with the exception that we add prepositions to the class of potential connectives, which are excluded by (M2) above. Our own computational lexicon of German connectives DiMLex (Stede, 2002) thus contains some prepositions (corresponding to the English *due to* and *despite*). Altogether, the XML-based DiMLex lists 170 frequent connectives and is used in both text generation and analysis applications.

On the issue of ambiguity, an investigation of the 170 DiMLex entries lead to the result that 42 also have non-connective readings.[2] For example, some of these cases are pronominal adverbs, which on the one hand can relate propositions and convey a coherence relation, and on the other hand can be used as event anaphors. One example:

– *Die Sprecherin verkündete das Ergebnis. Dabei half ihr ein Assistent.*
'The speaker announced the result. An assistant helped her with that.'

– *Die Sprecherin verkündete das Ergebnis. Dabei wollte sie eigentlich heute gar nichts sagen.*
'The speaker announced the result. Though she had planned not to say anything today.'

For human judges, the difficulty of deciding between the discourse/sentential reading varies to a great extent. For example, with *ferner* ('furthermore'/'farther away') it is easy, as the meanings of textual elaboration on the one hand, and relative spatial distance are quite distinct. With *dann* ('that way'/'in that case'/'thereafter'/'then'), which is always anaphoric, it can be quite difficult to say whether the antecedent is one specific point in time referred to by some portion of a previous clause, or the entire state of affairs expressed by the previous clause(s). Notice that the position of *dann* is not a clear cue for the disambiguation:

– *Wir können uns um fünf Uhr treffen. Dann bin ich im Cafe.*
'We can meet at five o'clock. I'll be in the cafe then.'

– *Wir können uns um fünf Uhr treffen. Dann haben wir eine Stunde Zeit für das Interview.*
'We can meet at five o'clock. That way / Thereafter we'll have one hour for the interview.'

For the experiments reported below in Section 4, we selected a subset of nine connectives, in order to test the methodology first on a small number of sample words. They are listed in Table 1 together with their part-of-speech (PoS) tags according to the STTS tagset (Schiller et al., 1999). To get an indication of the distribution of the non-/connective readings, we took a portion of 30,000 sentences from the TIGER corpus (Brants et al., 2004) and manually annotated the nine connectives, resulting in a Gold Standard for our experiments. In the frequency row of the table, the slash separates the PoS tags

[2] *aber allein allenfalls allerdings als also auch aufgrund außer da dabei dafür dagegen daher danach dann darauf darum denn doch entgegen ferner nebenher nur seitdem seit so sonst soweit statt trotz und während wegen weshalb weswegen wie wogegen womit wonach worauf zugleich*

Word	*allein*	*Allein*	*also*	*Also*	*auch*	*Auch*
STTS	ADV/ADJD	ADV	ADV	ADV	ADV	ADV
Frequency	141–1/1–0	26–2	3–99	1–9	768–19	92–17
Word	*dann*	*Dann*	*doch*	*Doch*	*ferner*	*Ferner*
STTS	ADV	ADV	ADV/KON	ADV/KON	ADV/ADJ	ADV/ADJ
Frequency	115–75	3–73	119–9/5–33	1–1/1–118	6–14/3–0	0–18/0–0
Word	*nur*	*Nur*	*so*	*So*	*sonst*	*Sonst*
STTS	ADV	ADV	ADV/KOUS	ADV/KOUS	ADV	ADV
Frequency	542–1	87–16	275–54/0–0	29–89/0–0	38–25	1–5

STTS tags: ADV = adverb, ADJ = adjective (ADJA: attributive ; ADJD: non-attributive), KON = coordinating conjunction, KOUS = subordinating conjunction

Table 1: Nine German connectives that also have non-connective readings

where applicable, and within these the dash separates the non-connective from the connective readings. The numbers show that in this particular corpus, some words are altogether rare (*ferner*), some do not or rarely occur as connectives (*allein*, *nur*) or almost always as connectives (*also*). To some extent, this reflects the genre of the language in the corpus (newspaper); for example, *allein* is used as a connective quite often in literary writing, but not in other genres. — Besides annotating this "real" data, we also hand-crafted a small test suite that contains two or three constructed sentences for each reading of the words, which are supposed to exemplify typical usages of the particular readings.

3 Robust discourse parsing

In the most common approach to robust text analysis nowadays, the first step is that of PoS tagging, which provides the foundation for subsequent analyses such as chunk parsing, rhetorical parsing, information extraction, and the like. For rhetorical parsing, i.e., the derivation of a possibly underspecified discourse structure, connectives are generally taken as the central (if not the only) source of information, cf. Corston-Oliver (1998), Marcu (2000), Schilder (2002), Hanneforth et al. (2003). It is therefore of great importance that connectives be *identified* correctly, in order to avoid adding more errors to an analysis task that is already very difficult.

With connectives being a syntactically heterogeneous class, however, no standard PoS tagger makes them readily available. Assuming the STTS tagset as the standard for German, we determined how state-of-the-art PoS taggers (TreeTagger (Schmid, 1994), TnT (Brants, 2000), Brill tagger trained on TIGER) deal with the nine words in question, using the 50 sentences of our hand-crafted test suite. It turns out that the taggers largely agree on their handling the words; differences occurred in 10% of the cases. In the following sections, to keep things simpler, we use only the results of the Brill tagger. Most of the nine words are constantly tagged as adverbs (regardless whether they function as connectives or not). *Doch* and *dann* can be adverb or conjunction, and when tagged as conjunction, this corresponds to the connective reading. Not all connective usages receive the 'KON' (= conjunction) tag, though. As for the other words, the non-/connective distinction cuts across the adverbial instances, and thus the most important aspect of our task is the disambiguation of an STTS 'ADV' tag.

4 Experiments: tagging connectives

In a series of experiments, we examined the use of common part-of-speech taggers to perform the disambiguation task, specifically: to assign a (non-standard) 'DC' tag to words used as discourse connectives. In this section, we present the experiments and results in detail.

The Brill tagger

In our experiments, we mainly focused on the Brill tagger, for two reasons: (i) during training, the Brill tagger acquires a set of symbolic rewriting rules, called *context rules*, many of which are plausible linguistic rules. After training, the context rules can be inspected and modified.[3] We present an example rule below. (ii) The Brill tagger allows for *incre-*

[3]See, e.g., Schneider and Volk (1998), who improved performance of a Brill tagger for German by manually adding context rules.

mental training; e.g., a tagger that has been trained on the STTS tagset can be trained further on a modified tagset, which includes tags for discourse connectives. This is an attractive option, since often large corpora annotated with standard tagsets are available for training. Hence, one goal is to investigate whether it is possible to successfully train a tagger on a large standard corpus combined with a small extra-annotated corpus.

Training the Brill tagger involves several steps. First, the tagger derives a lexicon from a fraction of the training corpus, which records for each word form its most frequent PoS tag. Next, the tagger acquires rules for guessing the PoS tags of unknown words. It first assigns pre-defined tags to capitalized and non-capitalized unknown words, which have to be specified manually. For German, we chose 'NN' (common noun) as the initial tag for capitalized words and 'ADJA' (attribute adjective) for non-capitalized words.[4] The tagger is then run on another fraction of the training corpus, which may include words hitherto unkown to the tagger. If the pre-defined tag for such an unknown word is incorrect, i.e., it does not correspond to the tag in the training corpus, the tagger evaluates variants of rewrite rules, which refer to prefixes or suffixes of the unknown word to determine its correct tag. Finally, the tagger derives context rules from another fraction of the training corpus: for each word which gets still tagged incorrectly, the tagger evaluates various forms of context rules, which refer to adjacent words and their tags to rewrite the incorrect tag. For instance, in our experiments, the tagger acquired the rule (R1), which can be paraphrased as "If a word has been initially tagged as 'KON' and the previous word is comma (= tagged with '$,'), then rewrite 'KON as 'DC' ".

(R1) `KON DC PREVTAG $,`

The training scenarios

For our experiments, we took 30,000 sentences from the TIGER corpus, together with their PoS annotations (STTS tags). 5,356 sentences (= 17.75%) contained at least one instance of the nine connectives ('DC') we were looking at. We manually went through the corpus and added tags marking connective vs. non-connective use. Since some of the DCs occurred very frequently (e.g., there were 2,059 instances of (non-capitalized) *auch*), we did not annotate all of the instances but 2,410 sentences only, with 2,938 DC instances. Table 1 lists these connectives, together with the frequencies of connective vs. non-connective use. In our experiments, we used this set of 2,410 sentences as the training and evaluation data.

In our scenario, we had to face the problem that we had annotated only a small selection of the class of connectives. Hence, it might turn out that the tagger would not learn to discriminate connectives successfully, as there is too much counter-evidence by the connectives we did not annotate and thus retain their original tag. However, we see our scenario as a realistic one: often, resource constraints put limitations on the amount of training data that can be manually annotated. We therefore thought it worthwhile to investigate the approach under these circumstances.

For the training scenarios, we defined two parameters that varied the amount of information that is encoded in the annotation. This resulted in four different scenarios. The parameters are:

(P1) Mark positive DC instances only — i.e., DCs used in the connective reading — vs. positive and negative ones.

(P2) Keep/do not keep record of the original STTS tag.

An example: Example (0) below displays the original TIGER STTS-annotation (format: *word/tag*) of the sentence fragment *Es geht also nicht nur um* ... ('Hence, it is not only about ...'). The fragment contains two adverbial connective candidates, *also* 'hence' and *nur* 'only', which are annotated with the STTS tag 'ADV' (adverb). *Also* indeed functions as a connective in this context, whereas *nur* does not have a connective reading here. (1) shows the input to training scenario 1, with positive marking only ('DC'). The non-connective *nur* keeps the original STTS tag. (2), used in scenario 2, also records the origial STTS tag, resulting in complex tags such as 'ADV_DC'. (3) marks positive and negative instances ('DC+/-'). Finally, (4) combines positive and negative marking with recording of the original tag.

[4] We determined these values as follows: we trained the TnT tagger on the German NEGRA corpus (Skut et al., 1998), which is annotated with STTS tags. Running this tagger on the German TIGER corpus revealed these tags as the most frequent ones of unknown words.

(0) Es/PPER geht/VVFIN **also/ADV** nicht/PTKNEG **nur/ADV** um/APPR ...

(1) Es/PPER geht/VVFIN **also/DC** nicht/PTKNEG **nur/ADV** um/APPR ...

(2) Es/PPER geht/VVFIN **also/ADV_DC** nicht/PTKNEG **nur/ADV** um/APPR ...

(3) Es/PPER geht/VVFIN **also/DC+** nicht/PTKNEG **nur/DC-** um/APPR ...

(4) Es/PPER geht/VVFIN **also/ADV_DC+** nicht/PTKNEG **nur/ADV_DC-** um/APPR ...

So, the scenarios differ in the specificity of the information added to the candidate words, which in turn leads to differences in the respective amounts of training data available for each scenario. For instance, moving from (3) to (4) increases the specificity of the DC tags, and accordingly the overall number of training instances for one particular DC tag decreases.

Adding a third parameter resulted in incremental variants of the scenarios 1–4: We first trained the Brill tagger on the STTS-annotated NEGRA corpus (Skut et al., 1998) and re-trained this model on our DC-annotation corpus.

Results

We performed a 4-fold cross-validation. On average, the training data consisted of 1,806 sentences, with 508 positive and 1,695 negative DC instances; evaluation data consisted of 602 sentences with 169 positive and 565 negative DCs.

As one kind of baseline, we trained and evaluated the Brill tagger on the original STTS-version of the sentences, without connective tags, as in example (0). We thus achieved 92.76% accuracy. As can be seen from the results displayed in Table 2, the DC-trained taggers perform comparably well, with respect to overall accuracy (first row).This baseline is important for making sure that adding our DC tags does not perturb the performance of the tagger in general, which would be a highly undesirable side effect.

Next, we computed baselines for DC annotation, by assigning to each DC candidate the tag it was assigned most often in the entire training corpus, given just its STTS tag. For instance, *doch* 'though' is ambiguous according to STTS and can be assigned 'ADV' or 'KON'. We found that in cases where it is analyzed as 'ADV', it most often is used in the sentential reading. In contrast, 'KON' signals connective use of *doch* (cf. Table 1). For computing the baseline, we used again the Brill tagger trained on the original STTS-version and then mapped the tags of DC candidates as described. These baselines, which can thus be characterized as the "STTS-tag majority view" are printed in italics in Table 2. The figures show that the STTS tags already encode a good amount of DC-relevant information, and so a static mapping from pairs of (DC candidate, STTS tag) to the set of {connective,non-connective} can solve the problem to a good extent. Notice that these mappings often have to be case-sensitive; for example, the rules for *doch* are:

```
doch  ADV  =>  DC-
doch  KON  =>  DC+
Doch  ADV  =>  DC+
Doch  KON  =>  DC+
```

Further, from the table we can deduce the following results:

- In all scenarios, training improved recall considerably. For instance, looking at the results of scenario 1, we see that the tagger now found 75.37% of connective use, as opposed to the baseline of 67.11%.[5]
- We also observe the well-known trade-off between precision and recall in all scenarios. That is, the tagger successfully learned to catch more of the positive instances, but it overgenerates.
- All in all, the figures for identifying positive instances seem rather low, wheras negative instances are tagged quite well. This reflects the fact that negative instances represent the "default" case.
- Among the four original scenarios, scenario 3 seems to be the overall winner, but it does not perform best in *each* of the prec/rec measures.
- The results from incremental training show that performance increases a little: overall accuracy ranges from 93.13%–93.16%. Recall of

[5] Note that in the scenarios with positive and negative marking, i.e., scenarios 3 and 4, the recall of *positive* DCs correlates with the precision of *negative* DCs (and likewise precision of positive DCs and recall of negative DCs). For instance, if a tag that has been incorrectly tagged as non-connective by the baseline tagger is now correctly analyzed as connective, the recall of positive DCs as well as the precision of negative DCs increase.

		1: DC$_+$ only	2: STTS+DC$_+$	3: DC$_{+/-}$	4: STTS+DC$_{+/-}$
Accuracy		**92.38%** 93.13% *92.32%*	**92.51%** 93.15% *92.31%*	**92.52%** 93.16% *92.35%*	**92.38%** 93.13% *92.31%*
DC$_+$	Precision	**83.91%** 81.72% *89.04%*	**79.69%** 84.88% *88.06%*	**84.46%** 85.84% *89.04%*	**83.53%** 85.29% *88.06%*
	Recall	**75.37%** 74.48% *67.11%*	**75.22%** 68.73% *66.37%*	**77.73%** 73.30% *67.11%*	**72.57%** 71.83% *66.37%*
	F-measure	**79.41%** 77.93% *76.53%*	**77.39%** 75.96% *75.69%*	**80.95%** 79.08% *76.53%*	**77.66%** 77.98% *75.69%*
DC$_-$	Precision	–	–	**93.47%** 92.33% *90.81%*	**92.17%** 92.02% *90.07%*
	Recall	–	–	**95.71%** 96.37% *97.52%*	**95.80%** 96.37% *96.73%*
	F-measure	–	–	**94.58%** 94.31% *94.05%*	**93.95%** 94.14% *93.28%*

Training scenarios: 1 = only positive instances marked; 2 = original STTS tag recorded; 3 = positive and negative instances marked; 4 = STTS and positive/negative marking. Results from the original scenarios are printed in boldface, those from incremental training in normal font. Baselines (according to the STTS-tag majority view) are printed in italics.

Table 2: Tagging results for the different scenarios

positive instances is not as good as in the original scenarios, the decrease being especially important with scenarios 2 and 3. Surprisingly, with scenario 1 the decrease of recall (compared to the original training scenario) does not correlate with an increase of precision. In general, though, the results of incremental training lie between the baselines and the corresponding original scenarios.

5 Summary and Outlook

We found that the standard PoS tags assigned to potential connectives by off-the-shelf taggers are not very reliable for connective identification, but that a fairly simple mapping (the baseline described in the last section) can improve the situation considerably. Going beyond this mapping with re-training improved the recall a lot, but pays the price of some reduction in precision. This overall finding is confirmed by a more qualitative analysis: the inspection of the results in our hand-crafted test suite of 50 sentences. The tagging results (using scenario 3) are only marginally better than the baseline, but it turns out that tagging and baseline have different strengths and weaknesses.

One obvious source of potential improvement is the amount of positive instances in the training data. For methodological reasons, we had decided to take 30,000 randomly-selected sentences from the TIGER corpus for DC-annotation; this has the effect that some of our words show up only very rarely as connectives (see Table 1). Adding more of these will probably make it easier for the tagger to acquire the discriminating features.

The other obvious limimitation is our restriction to only nine out of 42 words. This does not only mean that we have not learned about 33 words; in addition the tagging suffered from the fact that the training corpus contains counter-evidence of not-annotated DC candidates. Our next step will be to handle more words, but not on an individual basis, but by postulating equivalence classes in order to speed up the annotation process. For example, *nur* and *allein* can quite confidently be expected to behave very similar in the respects that we are interested in.

References

Diane Blakemore and Robyn Carston. 2005. The pragmatics of sentential coordination with 'and'. *Lingua*, 15(4):569–589.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Thorsten Brants. 2000. TnT — a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*.

Eric Brill. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT.

Simon Corston-Oliver. 1998. *Computing of Representations of the Structure of Written Discourse*. Ph.D. thesis, University of California, Santa Barbara.

Brigitte Grote. 2003. Signaling coherence relations in text generation: A case study of German temporal discourse markers. Doctoral dissertation, Universität Bremen.

Thomas Hanneforth, Silvan Heintze, and Manfred Stede. 2003. Rhetorical parsing with underspecification and forests. In *Proceedings of the HLT/NAACL Conference 2003*, Edmonton/AL.

Julia Hirschberg and Diane J. Litman. 1994. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Alistair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

Renate Pasch, Ursula Brauße, Eva Breindl, and Ulrich Hermann Waßner. 2003. *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. de Gruyter, Berlin.

Frank Schilder. 2002. Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering*, 8(2/3):235–255.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, University of Stuttgart and University of Tübingen.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Gerold Schneider and Martin Volk. 1998. Adding manual constraints and lexical look-up to a Brill-tagger for German. In *Proceedings of the ESSLLI-98 Workshop on Recent Advances in Corpus Annotation*, Saarbrücken.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of Ggerman newspaper texts. In *Proceedings of LREC-98*, pages 705–712, Granada.

Manfred Stede. 2002. DiMLex: A lexical approach to discourse markers. In A. Lenci and V. Di Tomaso, editors, *Exploring the Lexicon — Theory and Computation*. Alessandria (Italy): Edizioni dell'Orso.