Adding Semantic Relations to a Large-Coverage Connective Lexicon of German

Tatjana Scheffler and Manfred Stede

FSP Cognitive Science / Applied Computational Linguistics
University of Potsdam
Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany
tatjana.scheffler|manfred.stede@uni-potsdam.de

Abstract

DiMLex is a lexicon of German connectives that can be used for various language understanding purposes. We enhanced the coverage to 275 connectives, which we regard as covering all known German discourse connectives in current use. In this paper, we consider the task of adding the semantic relations that can be expressed by each connective. After discussing different approaches to retrieving semantic information, we settle on annotating each connective with senses from the new PDTB 3.0 sense hierarchy. We describe our new implementation in the extended DiMLex, which will be available for research purposes.

Keywords: connectives, lexicon, semantics

1. Introduction

Discourse connectives are closed-class lexical items that are known to provide very useful information for tasks like discourse parsing in the style of RST (e.g., (Hernault et al., 2010)) or PDTB (e.g., (Lin et al., 2014)), relation extraction (e.g., finding causal statements in biomedical text, (van der Horn et al., 2008)), or argumentation mining (e.g., (Peldszus and Stede, 2015)), because they indicate the kind of coherence relation holding between adjacent text spans (e.g., *because*: causal; *although*: concessive). While there can be a non-trivial ambiguity problem,[1] connectives are generally regarded as highly predictive cues. Having access to a list of connectives for a language, along with information on their syntactic and semantic/pragmatic behavior, is thus an important asset. In addition to the "straightforward" use in a discourse parser, connective lists can moreover help to generate training data for identifying unsignalled ('implicit') coherence relations, as shown by (Ji et al., 2015).

In this paper, we report on our recently completed extension of a German connective lexicon, which has grown both in the number of entries and in the amount of information per connective. In particular, we focus here on the difficult problem of selecting the coherence relations that are to be associated with lexical entries. First, Section 2. provides background on the lexicon, then Section 3. presents our current solution. We discuss results of the procedure in Section 4., and Section 5. concludes.

2. A Lexicon of Discourse Connectives

Our work extends DiMLex, the German Discourse Marker Lexicon (Stede and Umbach, 1998; Stede, 2002). We have now partially restructured the lexicon and added almost 100 new connectives. Our underlying definition of discourse connectives in German is based on (Pasch et al., 2003, p. 331):

(1) **Def.:** A *discourse connective* is a lexical item x that exhibits each of the following five properties:

(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 arguments of the relation (the meaning of x) are propositional structures.
(M5) The expressions of the arguments of the relation can be sentential structures.

Following (Stede, 2002), we drop M2 because our lexicon deliberately includes several prepositions that can be used as connectives (in the sense of M1, M3-M5), e.g., *trotz* ('despite') or *wegen* ('due to'). In German, these prepositions are often used with nominalized arguments that in every other way (semantically and pragmatically) resemble their sentential origins.

The lexicon now contains 275 German connectives that adhere to this definition and are in current use. Having compared the list to that of the extensive work of (Pasch et al., 2003) and used it for text annotation, we confirmed that our list is by and large complete. Current and future work addresses the structure and content of the individual entries. Currently, each entry in the publicly-available version of DiMLex[2] specifies:

- possible orthographic variants,
- ambiguity information (whether the lexical item also has non-connective readings),
- examples of non-connective readings,
- information on focus particles or correlates that can be associated with the connective, and
- syntactic category of the connective (different types of adverbs, conjunctions, and pre/postpositions are being distinguished).

[1] Some words have non-/connective uses; some connectives can signal more than one relation; see (Stede, 2014).

[2] https://github.com/discourse-lab/dimlex/

3. Assigning Discourse Relations

While the compilation of orthographic and morphosyntactic information for connectives is a relatively objective step, associating them with semantic/pragmatic discourse relations is methodologically far from trivial and prone to intersubjective disagreement. We see three central questions that need to be answered and that we will consider in turn:

1. Which set of relations is to be used?
2. On what grounds do we assign one or more relations to a specific connective?
3. How is ambiguity of different kinds represented?

3.1. Choice of relation set

Addressing the first question, there is a large range of discourse relation hierarchies that could potentially be used. Starting with (Knott, 1996), these schemas cover at least the four basic relation types (additive, conditional-causal, temporal, and contrastive), and provide different further granularities. The relation schemas from the PDTB (Webber et al., Submitted 2016), RST (Mann and Thompson, 1988), and SDRT (Asher and Lascarides, 2003) frameworks have received the most practical use for the annotation and analysis of discourse structure. These frameworks define an inventory of discourse relations of approximately similar granularity, and the inventories have considerable overlap (e.g., each contains a 'cause' relation). However, they are based on different foundations and approaches.

RST defines relations between text segments (elementary or complex discourse units), based on the cognitive effect of the combination on the reader. Though connectives may be present, and sometimes signal relations, they are not considered central for determining the identified relation. In addition, equal stretches of text can sometimes receive different relation assignments based on the larger discourse context (e.g., with regard to the nuclearity of the first or second segment).

SDRT is a semantic theory that assigns discourse relations as part of identifying a comprehensive semantic interpretation of the discourse. In as far as the lexical semantic content of discourse connectives is taken into account in the semantic computation, SDRT relations therefore can be mapped well onto connectives. On the other hand, several types of relations are missing from SDRT (e.g., negative ones like 'concession' (Roze et al., 2012)), and the strict split in coordinating and subordinating relations relies on a mix of syntactic and semantic properties that does not carry over in a clear way to connectives in our language.

Finally, the PDTB sense hierarchy (Webber et al., Submitted 2016) was developed from a lexicalized, flat representation of discourse structure. In PDTB, connectives are seen as the central vehicles for discourse relations (some other relations being 'implicit', i.e. unmarked), and each connective use is assigned one PDTB sense tag from the inventory. The new, improved PDTB 3.0 schema corresponds better to the other proposals (e.g., it includes some previously missing relations and exhibits a relatively flat hierarchy grouped in the four large classes mentioned above) and turns out to be quite useful for our purpose. Choosing the PDTB set of relations, one can see the list of senses associated with a particular connective in the lexicon as a list of *options*, i.e. the potential relations that this connective can signal in a text.

In summary, we decided to make PDTB senses the primary relational information in our lexicon. This means that the decision whether a connective is semantically ambiguous is made relative to the PDTB sense hierarchy (see Sct. 3.3.). At the same time, we wish to also store information that we can obtain from other sources, using other frameworks. Collecting these in the dictionary entries will enable systematic comparisons of the mis-/matches of the different frameworks – examples of which will be shown in Sct. 4.

3.2. Assigning relations to connectives

For obtaining information on which relation(s) to assign to a connective C, we see four different sources of information:

Annotated corpus: If there is an independently annotated corpus (for English: PDTB corpus, RST-DT, etc.), extract all discourse relations assigned to instances of C.

Lexicon in a different language: If there is a connective lexicon for some other language, map C to its translation correspondent(s) and extract the associated relation information from the target lexicon.

Grammars or other literature: Consult traditional linguistic or lexicographical resources about semantic/pragmatic information on C.

Intuitive judgement on raw corpus samples: Obtain samples of corpus instances for C and annotate them with relations (using annotation guidelines and substitution tests).

All these methods will by themselves lead only to partial information for our purposes, and the fourth in addition is costly. So, in order to gather information as comprehensively as possible, we used all four options in our work, restricting method 4 to a limited number of samples.

The **annotated corpus** at our disposal is the Potsdam Commentary Corpus (Stede and Neumann, 2014), which has been independently annotated for discourse connectives (without relations) and with RST discourse trees. We extracted each instance of a connective C and automatically associated it with the RST relation that corresponds to C in the text.[3] This way, we extracted between one and 19 attested possible relations for 126 connective types that occur in the PCC. About half of the connectives in DiMLex never occur in the PCC, due to the small size of the corpus and the infrequency of many connectives.

For an existing **lexicon**, we used LexConn (Roze et al., 2012), a large lexicon of French connectives. In joint work with Margot Colinet (Université Paris 7), we identified the correspondences between German and French connectives. We then transferred the LexConn relation annotations to the

[3] Not all connective instances could be matched, because some subsentential relations (such as center embeddings) were not annotated in the RST structures and there were issues with non-binary-branching relations.

corresponding German connective. LexConn uses a relation set based on SDRT, but with some relation types added from the RST inventory. This method yields relation annotations for 149 of the connectives in DiMLex. Of course, connectives of syntactic types that have no equivalent in the French lexicon (such as prepositions) cannot receive relation assignments in this way.

Traditional **linguistic grammars** are a valuable resource especially for well-researched languages like German. We collected connective senses from the (Helbig and Buscha, 1984) grammar of German. Like most grammers, it does not use 'connective' as a unified category, though, so the information had to be collected from portions on adverbials, conjunctions, and prepositions, and non-connective readings had to be filtered. It turns out that half of the DiMLex connectives (mostly rare and/or phrasal items) do not appear in that grammar and thus do not receive relational information through this method. Further, Helbig and Buscha use their own semantic classification schema that differs slightly from the ones we discussed in Sct. 3.1.

Finally, we collected 100 **raw corpus samples** for each of our connectives from the DWDS[4] corpus collection of newspaper text (Geyken, 2014). We applied the new PDTB 3.0 annotation manual (see Sct. 4.2. for more details). This method guarantees complete coverage of the entries in our lexicon. Many connectives (like *außerdem* 'in addition') are quite straightforward and unambiguous, but ambiguous (*während* 'while') or vague (*aber* 'but, though, however') connectives may not exhibit all usage variants in a small number of instances. Hence, there is no guarantee that all semantic readings occur in the sample set, so that the information from methods 1-3 is needed to supplement our final assignment of relations in the lexicon entries.

3.3. Representation in the lexicon

We now explain the design decisions for our new extension of the lexicon, which is encoded in XML. The first principle is to retain a primary division of lexical entries into syntactic frames. That is, when a connective has more than one syntactic category, each of them is stored in a separate `<syn>` section, where the appropriate syntactic features are collected.

Within each syntactic category, as explained above, PDTB senses are taken to motivate the semantic reading(s). Thus for every syntactic category we decide whether more than one PDTB sense can be assigned; each of these generates a `<sem>` field within the corresponding `<syn>` field. Consider the example of *während*, which corresponds to English 'while' and 'during'. It has two `<syn>` fields, the first for the preposition, the second for the subordinator. When *während* is used as a preposition, it can only have a temporal reading; therefore there is just one `<sem>` field describing the PDTB sense 'synchronous' (and possibly further semantic information such as role linking and examples, which will be added in future work). When used as a subordinator, this connective has the same ambiguity as English 'while' and thus receives two `sem` fields within the `syn`, one for 'contrast' and one for 'synchronous'. — The drawback of this approach is that we need to duplicate the semantic information in case there is syntactic ambiguity but no semantic one, but this seems not too dramatic, given the gain in overall transparency. Figure 1 shows a simplified representation of the lexical entry for *während*.

The remaining question is where to store the information from the other sources. Since our lexicon at this point does

```
<entry id="k173" word="während">
    <dict_info>
      <sdrt_lexconn>
        <concession/>
        <background/>
        <contrast/>
        <temploc/>
      </sdrt_lexconn>
      <helbig_buscha>
        <temporal>
          <gleichzeitigkeit/>
        </temporal>
        <adversativ/>
      </helbig_buscha>
      <rst>
        <antithesis/>
        <contrast/>
        <circumstance/>
      </rst>
    </dict_info>
    <syn>
      <cat>subj</cat>
      <sem>
        <coherence_relations>
          <synchronous />
          <contrast />
        </coherence_relations>
      </sem>
      <sem>
      </sem>
    </syn>
    <syn>
      <cat>praep</cat>
      <praep>
        <ante>1</ante>
        <post>0</post>
        <circum>0</circum>
        <case>gen</case>
      </praep>
      <sem>
        <coherence_relations>
          <synchronous />
        </coherence_relations>
      </sem>
    </syn>
  </entry>
```

Figure 1: Simplified lexical entry of *während*.

[4] www.dwds.de

connective	SDRT/LexConn	PDTB 3.0	RST/PCC	Helbig/Buscha[5]
aber	opposition	concession-arg1-as-denier concession-arg2-as-denier contrast conjunction	concession antithesis background list/joint	adversative
außerdem	continuation	conjunction	list	conjunctive
während	concession background contrast temploc	contrast synchronous	antithesis contrast circumstance	adversative synchronicity
weil	explanation explanation*	cause-reason	cause reason	causal

Table 1: Selected connectives and their associated relations.

not intend to make any commitment on the mapping between different relation hierarchies (which is an unresolved research issue), we do not relate the other senses to the PDTB ones, but instead store them on the top level of the lexicon entry in a `dictionary-info` field, which has subfields for RST, SDRT/LexConn, and Helbig/Buscha. In this way, all the information about a connective is assembled within the lexicon entry and can be exploited for studying correlations between the different sense/relation inventories.

4. Results

In our view, collecting different semantic analyses through these different methods results in a richer and ultimately more useful resource, for at least two reasons. First, if two analyses are found to agree or can be mapped to one another, this increases their trustworthiness, since different kinds of evidence have contributed to the finding. Second, in cases where two analyses in different frameworks do not agree, the differences often point out new and interesting research problems rather than mere errors.

4.1. Case Studies

In Table 1 we show the semantic relations we obtain for four frequent connectives. These examples illustrate our approach. Some connectives are semantically quite straightforward. For example, *weil* ('because') is the most common causal subordinating conjunction in German. Each framework provides a basic causal relation, which is the one commonly assigned to *weil*.[6] Some frameworks provide additional granularity wrt. volition (RST) or epistemicity (SDRT), but one can easily map these facets to a basic causal sense. Similarly, *außerdem* ('in addition') marks an additive discourse relation in each formalism. These connectives illustrate the first case above, where converging analyses reassure the researcher of their validity.

Some other cases are more complex. For example, *aber* ('but') is tagged as merely contrastive in the traditional German grammar, as well as in the SDRT relations obtained through LexConn ("opposition"). The corpus annotations in RST and PTDB, however, also indicate a narrower concessive reading of the connective. In addition, both methods also identified a non-contrastive additive reading ("conjunction" in PDTB, "list/joint" in RST) closer to *but also* that might have been missed in the other work. Here, the multi-pronged approach helped validate the analyses and identify missing relations.

Finally, the subordinating *während* ('while') is ambiguous between a temporal ("synchronous") and a contrastive reading (both readings are about equally frequent). This ambiguity is also confirmed by the lexical work in (Helbig and Buscha, 1984). In the SDRT correspondences, we can observe the different focus of the relation definitions: Most temporal instances of 'while' were analysed as "background", an additive relation that is used in order to provide additional, but less important (subordinated) information about a situation. The analysis focuses more on the relative status of the information (subordinated or not) rather than the temporal relation (at the same time). This difference in focus might be facilitated by the SDRT framework itself. In RST, the temporal reading similarly corresponds to "circumstance". However, it is not at face value clear in which way *während* is contrastive: Does it express a mere (same-level) contrast, or is it concessive by indicating one argument as more central than the other? In the LexConn version of SDRT, for whom the subordinating/coordinating distinction is central, the subordinator *während* receives a "concessive" tag in most cases (note that the relation "concessive" was added to the SDRT inventory by (Roze et al., 2012)). The RST corpus also identifies an "antithesis" use which is a nucleus/satellite (subordinating) relation. However, according to the (semantic) PDTB definition of concession (a denial of expectation), we could not find any concessive uses of *während* in 100 instances from German newspaper text, leaving only general "contrast". This opens up a discourse syntactic/semantic question about the definition of "concession" that is beyond the scope of this lexical work.

[5] Helbig/Buscha's German relation names have been translated: 'adversativ' = 'adversative', 'kopulativ' = 'conjunctive', 'Gleichzeitigkeit' = 'synchronicity', 'kausal' = 'causal'.

[6] Of course, *weil* can also be used for pragmatic justifications, either of a belief or a speech act. Even though PDTB 3.0 provides distinct senses for "cause+belief" and "cause+speech act", it is argued that these additional facets might be better accounted for by using features on top of some discourse connective senses. In fact, pragmatic uses are also known for concessives, conditionals, and other relations. Therefore, we only assign a basic overarching "cause" relation here for PDTB.

4.2. Agreement

We have tested the applicability of the new PDTB 3.0 annotation schema to our German newspaper data with two independent expert annotators (the authors), in order to obtain first agreement values. (Another study with guideline-trained annotators will follow later.) We picked five previously unannotated connectives of different syntactic types and complexities:

falls ('in case'), a subordinator with a conditional meaning,

ferner ('furthermore'), an adverbial with additive semantics,

folglich ('consequently'), an adverbial that has been characterized as marking the result clause of a causal or purpose relation,

freilich ('admittedly, indeed, however'), an adverbial with many different uses,

gleichzeitig ('at the same time'), an ambiguous adverbial similar to 'while' with a temporal and a contrastive sense.

The two annotators classified 50 instances for each connective according to the PDTB 3.0 schema (27 senses) with the additional option of marking an instance as not being a connective. This option was chosen when the lexical item occurred in a non-connective sense (e.g. *ferner* can also be used as the comparative adjective 'further (from)'), or when the context did not suffice to decide whether the instance was used as a connective (and in which sense). The annotation was carried out using WebAnno (Yimam et al., 2013) and the agreement computed with R.

connective	agree	Fleiss' κ	Jaccard
falls	95.7	-0.02	1
ferner	93.9	0.85	1
folglich	72.7	-0.02	0.17
freilich	42.9	0.20	0.86
gleichzeitig	61.4	0.41	0.60
all	75.2	0.71	0.72 (mean)

Table 2: Annotation agreement between 2 annotators.

Table 2 shows the agreement for the five connectives. The raw agreement values confirm the intuitive rating of the difficulty of the connectives: for the unambiguous connectives *falls* and *ferner*, the raw agreement is very high. Kappa values suffer from the high expected agreement in these cases, since only few senses are taken into account, and are not reliable for these connectives (third column). In addition, we computed the overlap between the sets of senses ever assigned by the two annotators (fourth column, Jaccard index). Since our ultimate goal is not a corpus of annotated instances, but an enriched lexicon, we only need to obtain a set of senses that each connective can express. Thus, a high Jaccard index[7] indicates that the sets of senses chosen by the annotators for this connective overlap significantly, and therefore that the sense annotation in the lexicon is robust. *Freilich* illustrates a class of highly ambiguous connectives. The raw agreement as well as chance corrected agreement for the sense annotation of individual instances is quite low (second and third column), because the meaning of the connective is vague and hard to differentiate in each instance. The high overlap of assigned senses between the two annotators (fourth column), however, shows that lexicographically, it is straightforward to determine the kinds of senses this connective can express[8].

Finally, the low agreement on *folglich* and especially *gleichzeitig* indicated that these need additional work in adjudication: For both connectives, one annotator assigned additional senses that the other annotator did not deem applicable. For example, one annotator identified several "contrastive" or "concessive" uses of *gleichzeitig* ('at the same time'), where the other annotator saw the basic temporal synchronicity sense as sufficient. Therefore, we discussed all connectives with the entire team of three annotators and decided difficult cases by consensus.

5. Conclusion

We presented our extension to the German DiMLex lexical resource of discourse connectives, which now has wide coverage, and in particular, all the entries have been assigned semantic/pragmatic relational information. We discussed different means of obtaining such semantic information, as well as our way of representing various kinds of ambiguities in the lexicon. Finding the "right" relations is by no means simple, and we consider it important to assemble information from different sources in the lexicon entries; this now enables systematic comparisons. At the same time, when a discourse parser (or similar software) is to make use of DiMLex, it is important to also have one single layer of sense information; we argued that the new PDTB 3.0 sense hierarchy is a good choice for this task, and extended our entries accordingly. The lexicon in the form described here will be freely available for research purposes at its Github location `https://github.com/discourse-lab/dimlex` when it is completed.

6. Acknowledgements

Thanks to Erik Haegert for his help with collecting information for DiMLex and building lexicon entries. We are grateful to the research group *Deutsches Textarchiv* at Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) for their assistance with obtaining corpus examples for our connectives. Our collaboration with Margot Colinet was supported by the European COST action TextLink through a Short Term Scientific Mission.

7. References

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

Alexander Geyken. 2014. Methoden bei der Wörterbuchplanung in Zeiten der Internetlexikographie. *Lexicographica*, 30(1):77–111.

[7] The Jaccard index between sets A and B is defined as $|A \cap B|/|A \cup B|$.

[8] One annotator also saw one instance of mere "conjunction", which the other annotator never assigned.

Gerhard Helbig and Joachim Buscha. 1984. *Deutsche Grammatik*. Verlag Enzyklopädie.

Hugo Hernault, Hemut Prendinger, David duVerle, and Mitsuru Ishizuka. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2224, Lisbon, Portugal, September. Association for Computational Linguistics.

Alistair Knott. 1996. *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

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

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

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal, September. Association for Computational Linguistics.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: A french lexicon of discourse connectives. *Discours [En ligne]*, 10. http://discours.revues.org/8645.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Manfred Stede and Carla Umbach. 1998. Dimlex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1238–1242.

Manfred Stede. 2002. DiMLex: A lexical approach to discourse markers. In *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria.

Manfred Stede. 2014. Resolving connective ambiguity: A prerequisite for discourse parsing. In *The Pragmatics of Discourse Coherence*. John Benjamins, Amsterdam.

Pieter van der Horn, Bart Bakker, Gijs Geleijnse, Jan Korst, and Sergei Kurkin. 2008. Determining causal and non-causal relationships in biomedical text by classifying verbs using a naive bayesian classifier. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 112–113, Columbus, Ohio, June. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. Submitted, 2016. Extending the Penn Discourse Treebank: Discourse relations and conjoined VPs. 10th Linguistic Annotation Workshop (LAW-X), Berlin.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Stroudsburg, PA, USA, August. Association for Computational Linguistics.