Constructing a Lexicon of English Discourse Connectives

Debopam Das and **Tatjana Scheffler** and **Peter Bourgonje** and **Manfred Stede**
Applied Computational Linguistics
UFS Cognitive Sciences
University of Potsdam / Germany
firstname.lastname@uni-potsdam.de

Abstract

We present a new lexicon of English discourse connectives called DiMLex-Eng, built by merging information from two annotated corpora and an additional list of relation signals from the literature. The format follows the German connective lexicon DiMLex, which provides a cross-linguistically applicable XML schema. DiMLex-Eng contains 149 English connectives, and gives information on syntactic categories, discourse semantics and non-connective uses (if any). We report on the development steps and discuss design decisions encountered in the lexicon expansion phase. The resource is freely available for use in studies of discourse structure and computational applications.

1 Introduction

Discourse connectives are generally considered to be the most reliable signals of coherence relations, and they are widely used in a variety of NLP tasks involving the processing of coherence relations, such as discourse parsing (Hernault et al., 2010; Lin et al., 2014), machine translation (Meyer et al., 2011), text summarization (Alemany, 2005), or argumentation mining (Kirschner et al., 2015). Accordingly, corpora annotated for discourse connectives and coherence relations have been developed for different languages.

In addition to discourse-annotated corpora, a *lexicon* of discourse connectives, giving the list of connectives for a language, along with useful information about their syntactic and semantic-pragmatic properties, can also serve as a valuable resource. Such lexicons were developed and are becoming more and more available in different languages, beginning with German (Stede and Umbach, 1998), later for Spanish (Briz et al., 2008) and French (Roze et al., 2010), and more recently for Italian (Feltracco et al., 2016), Portuguese (Mendes et al., 2018) and Czech (Mírovský et al., 2017).

We present a lexicon of English discourse connectives called DiMLex-Eng, which is developed as a part of the Connective-Lex database[1] at the University of Potsdam. It includes 149 connectives, a large part of which was compiled from the annotations of the Penn Discourse Treebank 2.0 (Prasad et al., 2008). We expanded that list to include additional connectives from the RST Signalling Corpus (Das et al., 2015) and relational indicators from a list supplied by Biran and Rambow (2011). For organizing the entries in the lexicon, we use the format of DiMLex, a lexicon of German connectives (Stede and Umbach, 1998; Scheffler and Stede, 2016). For each entry in DiMLex-Eng, we provide information on the possible orthographic variants of the connective, its syntactic category, non-connective usage (if any), and the set of discourse relations indicated by the connective (with examples from corpora). We describe our criteria for filtering connective candidates for inclusion in the lexicon, and give an outlook on the relationship between connectives and the broader range of 'cue phrases' or 'AltLex' expressions in language.

2 Sources of English connectives

2.1 The PDTB corpus connective list

The Penn Discourse Treebank corpus (PDTB, Prasad et al., 2008) is the best-known resource for obtaining English connectives. In the PDTB, connectives are defined as discourse-level predicates that take as their arguments two abstract objects such as events, states, and propositions, and that

[1] http://connective-lex.info/

Proceedings of the SIGDIAL 2018 Conference, pages 360–365,
Melbourne, Australia, 12-14 July 2018.

number	category	number	category
67	ADVP	2	NN
25	phrase	2	JJ
20	IN	2	INTJ
26	PP	1	VB
12	RB	1	RBR
8	CC	1	NNP
2	UCP	1	WHNP

Table 1: Distribution of syntactic types for connectives in the PDTB.[4]

are generally expressible as clauses.[2] In addition to *explicit* connectives, the PDTB contains *implicit* connectives: In the absence of an explicit connective, annotators insert an extra one that best signals a relation between two discourse segments. The PDTB also provides annotations of *AltLex* (alternative lexicalization) for instances where adding an implicit connective would lead to a redundancy in expressing the relation, since it is already conveyed by an indicative phrase.

The PDTB annotators were given the above-mentioned definition of 'connective' and asked to identify words/phrases that accord to this definition. In the end, 100 distinct connectives were marked in the corpus. This list of words was later routinely used by researchers working on shallow discourse parsing in order to find connective candidates in text. However, since the list of connectives is based on annotations of a particular corpus, no claim of exhaustivity of this list was ever raised. Since the corpus is annotated with parse trees and sense relations, the distribution of syntactic types and semantic relations attested for each connective can also be extracted. Table 1 shows the overall distribution of syntactic types for the connectives in the PDTB (note that one connective can have several syntactic types).

2.2 The RST Signalling Corpus

The RST Discourse Treebank (RST-DT, Carlson et al., 2003) is the largest and most widely-used corpus for developing discourse parsers for the framework of Rhetorical Structure Theory (Mann and Thompson, 1988). In contrast to the PDTB, it does not contain any markup of connectives; rather, it is restricted to representing the coherence relations among text segments. Recently, however, the RST-DT has been enriched with markup on *relation signals* in the RST Signalling Corpus (Das et al., 2015) (henceforth RST-SC): Going through every coherence relation in the corpus manually, the authors decided for each what signal (if any) can be located in either of the two related spans, which would aid the reader in identifying the relation. This goal leads to marking not only connectives, but also other lexical, semantic, syntactic, layout, or genre-based features. In the RST-SC, about 18 percent of all the relations are indicated by connectives or other discourse markers, which are distributed over 201 different types.

2.3 RST-DT relational indicator list

Also aiming at identifying lexical signals of relations, Biran and Rambow (2011) used a semi-automatic approach: They extracted all instances of relations (i.e., pairs of two text spans) from the RST-DT, and automatically identified the most indicative (1..4)-grams of words using a variant of TF/IDF. The n-grams were ranked, and an empirically-determined cutoff demarcated the list. The authors were specifically interested in argumentative relations and thus added a manual filtering step for a relevant subset of RST relations. However, they made a list of 230 indicators for all relations available.[5] The indicators range from one to four-word expressions, many of which qualify as discourse connectives: conjunctions (*but, although)*, prepositional phrases (*for instance, in addition*) or adverbials (*probably*).

The list also contains items belonging to different lexical categories, such as nouns (*statement, result*), verbs (*concluded, to ensure*) or other elements which simply comprise random strings of words and do not neatly represent any syntactic constituents (e.g., *and we certainly do, and just as we*). These items would be rejected as discourse connectives by any definition from the literature, and the procedure was of course not meant to result in a list of connectives per se. Yet, using this procedure, one could expect to also find quite a few proper connectives. As an explanation of why their number is, however, relatively

[2] In some exceptional cases, the arguments in the PDTB can also be realized as non-clausal structures, such as VP coordinates, nominalizations, or anaphoric expressions representing abstract objects.

[4] 'Phrase' indicates that the connective consists of more than one partial tree; otherwise, the single category that dominates the entire connective was chosen.

[5] http://www.cs.columbia.edu/~orb/code_data.html

small, note that relations are often realised without any explicit connective, thus lowering their co-occurrence numbers. Additionally, since a connective can be ambiguous in terms of the senses it represents, its distribution relative to one particular sense is less pronounced when it also accompanies other senses.

3 DiMLex

We chose to develop DiMLex-Eng using the format of the German DiMLex (DIscourse Marker LEXicon).[6] Its current version (Scheffler and Stede, 2016) contains an exhaustive list of 275 German discourse connectives. Following Pasch et al. (2003), (with one modification to be discussed in the next section), a connective in DiMLex is defined as a lexical item x which has the following properties: (i) x cannot be inflected; (ii) the meaning of x is a two-place relation; (iii) the arguments of this relation are propositional structures; (iv) the arguments can be expressed as sentential structures. This definition is comparable to the one used in the PDTB. Both frameworks consider a connective as a relational signal taking two semantic arguments.

For each entry, DiMLex provides a number of features, characterizing its syntactic, semantic and pragmatic behaviour. DiMLex has recently been incorporated in the Connective-Lex database (see Section 1), developed as part of the European COST action TextLink[7], and DiMLex-Eng is being included there as well.

4 Merging the sources into DiMLex-Eng

Our selection of entries in DiMLex-Eng follows from what we consider as English discourse connectives. The definition is partly based on that used for German connectives in DiMLex (provided in Section 3), and further modified by incorporating some features from the annotation in the PDTB. We consider a word or phrase x as a connective in English if it has the following properties:

- x cannot be inflected.
- The meaning of x is a two-place relation.
- The arguments of this relation are abstract objects (propositions, events, states, or processes).
- Usually, the arguments are expressed as clausal or sentential structures. However, they can also be expressed by phrasal structures (e.g., noun phrases beginning with connectives like *according to*, *because of*, or *given*) as long as they denote abstract objects.

Furthermore, we used the following two lexicographic exclusion criteria to determine whether a connecting phrase x which signals a coherence relation (as defined above) warrants inclusion in the lexicon as a connective entry:

1. x should be a fixed expression and cannot be freely modified by inserting other material.
2. x is not semantically compositional with respect to its component parts.

Criterion 1 excludes free phrases such as *for this reason* which can be modified: *for this excellent reason*, *for these reasons*, etc. Criterion 2 excludes phrases which consist of a connective and an intensifier/adverb such as *particularly if* or *especially when* (here, only *if* and *when* are considered connectives with their own lexicon entries), and also items comprising two connectives such as *and therefore* or *but at the same time*. According to this criterion, however, phrases such as *even though* and *even if* are considered to be distinct connectives, since their meaning is not straightforwardly compositional.

Once we decided on the definition of English connectives, we began compiling the lexicon with entries from the PDTB 2.0. We decided to include all 100 explicit connectives from the corpus, because they adequately fulfill our definitional requirements for connectives.

In the lexicon expansion phase, we first added more connectives from the RST-SC (Das et al., 2015). We observed that of the 100 PDTB connectives included in the initial version of DiMLex-Eng, 71 connectives are also found in the RST-SC, adding up to 3.390 instances (of marker tokens or phrases). More importantly, in the opposite direction, from the RST-SC, we added 46 connectives (which do not occur in the PDTB) to DiMLex-Eng. The resulting 146 entries cover 3.721 instances in the RST-SC (an extra 331 compared to the initial version of DiMLex-Eng). The RST-SC contains 201 types (3.899 instances). Note that we add only a subset of these to DiMLex-Eng due to the restrictions on entries explained above. With our extended lexicon, we now cover 117 of 201

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

[7] http://www.textlink.ii.metu.edu.tr/connective-lex

types (58%) and 3.721 of 3.899 instances (95%), compared to 35% (types) and 87% (instances) for the initial lexicon version that included just the PDTB-based list.

In the final phase of entry collection, we consulted the relational indicator list of Biran and Rambow (2011), and screened out only those items which satisfy our definition of discourse connective. We found that of the 230 entries in the Biran and Rambow list, seven items overlap with our 44 entries already selected from the RST-SC. Additionally, 12 of the 230 items were in the list initially extracted from the PDTB 2.0. Upon manual evaluation of the remaining 211 entries, we found five more connectives that we added to our lexicon.

5 Populating the lexicon entries

DiMLex-Eng includes significant lexicographic information about the syntactic and semantic-pragmatic properties of connectives. For syntactic and other non-discourse features of a connective entry, it specifies: (i) possible orthographic variants, (ii) ambiguity information (whether the lexical item also has non-connective readings), (iii) the syntactic category of the connective (see Table 1; mainly: adverb, subordinating conjunction/preposition, coordinating conjunction, or phrase), (iv) possible coherence relations expressed by the connective, (v) examples[8] of relations associated with the connective .

The semantic information about coherence relations was derived from the observed corpus instances in the cases of connectives from the PDTB and RST-SC. That is, each entry lists all coherence relations with which the connective occurred, together with frequency information.

For encoding the lexicographic features in DiMLex-Eng, we use the format of DiMLex, which provides a cross-linguistically applicable XML schema. Figure 1 shows a representation of the lexical entry for *by contrast* in DiMLex-Eng. The entry shows that *by contrast* is a PP which can be used to signal three possible coherence relations: CONTRAST (occurring 11 out of 27 times when *by contrast* was used as a connective in the corpus), JUXTAPOSITION (12 times), and OPPOSITION (4). The lexicon is being extended with corpus examples for each sense, where available.

[8]Mostly taken from the PDTB, RST-SC and Corpus of Contemporary American English (https://corpus.byu.edu/coca/)

6 Summary and Outlook

We have presented DiMLex-Eng, a lexicon of English discourse connectives, compiled from annotated corpora and modeled after DiMLex, a lexicon of German discourse connectives. The connectives in DiMLex-Eng are lexically frozen expressions (e.g., *because, furthermore, since*) that correspond to what are described by Danlos et al. (2018) as *primary* connectives (with respect to their degree of grammaticalization). The knowledge of such connectives along with their manually curated syntactic and discourse attributes, as the one offered by DiMLex-Eng, are valuable in areas such as language learning and contrastive discourse studies. Also, the connectives in DiMLex-Eng, together with other coherence relation signals, can serve as a valuable resource for discourse parsing and related applications.

Coherence relation signals, not necessarily restricted to being discourse connectives, may also comprise many other items, which are discussed under the labels of *cue phrase* (Knott and Dale, 1994), *secondary connective* (Danlos et al., 2018), *AltLex* expression (Prasad et al., 2008), or *relational indicator* (Biran and Rambow, 2011). These are more difficult to describe systematically and hence are less amenable to a lexical treatment; we leave it to future work to extend DiMLex-Eng into this direction.

We would like to point out that using the approach of selecting words and phrases that frequently co-occur with coherence relations, we find only 24 words or phrases that fulfill the constraints of true (primary) connectives, compared to the complete lexicon of 149 entries. This seems to imply that simple statistical co-occurrence measures are not sufficient for identifying discourse connectives, which must satisfy syntactic and semantic criteria, as well.

Another approach for automatic generation of discourse connective lexicons is by translational mapping between parallel corpora, which we are pursuing in ongoing work (Bourgonje et al., 2017), following up on earlier studies such as that of Cartoni et al. (2013). We hope to use this approach to identify additional connectives for DiMLex-Eng as well as establish and enhance correspondences between DiMLex-Eng and other similar connective lexicons.

```
<entry id="67" word="by contrast">
   <orths>
      <orth canonical="0" orth_id="67o1" type="cont">
         <part type="phrasal">By contrast</part>
      </orth>
      <orth canonical="1" orth_id="67o2" type="cont">
         <part type="phrasal">by contrast</part>
      </orth>
   </orths>
   <syn>
      <cat>PP</cat>
      <sem>
         <pdtb2_relation anno_N="27" freq="11"
                         sense="Comparison.Contrast" />
      </sem>
      <sem>
         <pdtb2_relation anno_N="27" freq="12"
                         sense="Comparison.Contrast.Juxtaposition" />
      </sem>
      <sem>
         <pdtb2_relation anno_N="27" freq="4"
                         sense="Comparison.Contrast.Opposition" />
      </sem>
   </syn>
</entry>
```

Figure 1: DiMLex-Eng entry for the connective *by contrast*.

Acknowledgments

Our work was financially supported by Deutsche Forschungsgemeinschaft (DFG), as part of (i) project A03 in the Collaborative Research Center 1287 "Limits of Variability in Language" and (ii) project "Anaphoricity in Connectives".

References

Laura Alonso i Alemany. 2005. *Representing discourse for automatic text summarization via shallow NLP techniques*. PhD dissertation, Universitat de Barcelona.

Or Biran and Owen Rambow. 2011. Dentifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 5(4):363–381.

Peter Bourgonje, Yulia Grishina, and Manfred Stede. 2017. Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics*, Rome, Italy.

Antonio Briz, Salvador Pons Bordería, and José Portolés. 2008. Diccionario de partículas discursivas del español.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer, Dordrecht.

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. 2013. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. 4:65–86.

Laurence Danlos, Kateřina Rysová, Magdaléna Rysová, and Manfred Stede. 2018. Primary and secondary discourse connectives: definitions and lexicons. *Dialogue and Discourse*. To appear.

Debopam Das, Maite Taboada, and Paul McFetridge. 2015. RST Signalling Corpus, LDC2015T10. Linguistic Data Consortium.

Anna Feltracco, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. 2016. LICO: A Lexicon of Italian Connectives. In *Proceedings of the Second Italian Conference on Computational Linguistic (CLiC-it 2016)*.

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.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2015 NAACL-HLT Conference*.

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 20(2):151184.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Amalia Mendes, Iria del Rio, Manfred Stede, and Felix Dombek. 2018. A Lexicon of Discourse Markers for Portuguese: LDM-PTs. In *Proceedings of the 11h International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *Proceedings of the SIGDIAL 2011 Conference*, SIGDIAL '11, pages 194–203, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiří Mírovský, Pavlína Synková, Magdaléna Rysová, and Lucie Políková. 2017. CzeDLex A Lexicon of Czech Discourse Connectives. In *the Prague Bulletin of Mathematical Linguistics*, volume 109, pages 61–91.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrackech, Morocco.

Charlotte Roze, Danlos Laurence, and Philippe Muller. 2010. LEXCONN: a French Lexicon of Discourse Connectives. In *Multidisciplinary Approaches to Discourse - MAD 2010*, Moissac, France.

Tatjana Scheffler and Manfred Stede. 2016. Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

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*. Association for Computational Linguistics.