Discours
Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

24 | 2019
Varia

Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives

Manfred Stede, Tatjana Scheffler and Amália Mendes

Electronic version
URL: http://journals.openedition.org/discours/10098
DOI: 10.4000/discours.10098
ISSN: 1963-1723

Publisher:
Laboratoire LATTICE, Presses universitaires de Caen

Electronic reference
Manfred Stede, Tatjana Scheffler and Amália Mendes, « Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives », *Discours* [Online], 24 | 2019, Online since 30 October 2019, connection on 01 November 2019. URL : http://journals.openedition.org/discours/10098 ; DOI : 10.4000/discours.10098

discours
Revue de linguistique, psycholinguistique et informatique
http://journals.openedition.org/discours/

Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives

Manfred Stede
Potsdam University

Tatjana Scheffler
Potsdam University

Amália Mendes
CLUL (Centro de Linguística da Universidade de Lisboa)
Lisbon University

Manfred Stede, Tatjana Scheffler, Amália Mendes, « Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives », *Discours* [En ligne], 24 | 2019, mis en ligne le 30 octobre 2019.

URL : http://journals.openedition.org/discours/10098

Titre du numéro : *Varia*
Coordination : Anne Le Draoulec & Josette Rebeyrolle

Date de réception de l'article : 07/01/2019
Date d'acceptation de l'article : 29/03/2019

Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives

Manfred Stede

Potsdam University

Tatjana Scheffler

Potsdam University

Amália Mendes

CLUL (Centro de Linguística da Universidade de Lisboa)
Lisbon University

In this paper, we present a tangible outcome of the TextLink network: a joint online database project displaying and linking existing and newly-created lexicons of discourse connectives in multiple languages. We discuss the definition and demarcation of the class of connectives that should be included in such a resource, and present the syntactic, semantic/pragmatic, and lexicographic information we collected. Further, the technical implementation of the database and the search functionality are presented. We discuss how the multilingual integration of several connective lexicons provides added value for linguistic researchers and other users interested in connectives, by allowing crosslinguistic comparison and a direct linking between discourse relational devices in different languages. Finally, we provide pointers for possible future extensions both in breadth (i.e., by adding lexicons for additional languages) and depth (by extending the information provided for each connective item and by strengthening the crosslinguistic links).

Keywords: discourse connectives, lexicon, multilingual resources, crosslinguistic links

Nous présentons dans cet article un résultat tangible du réseau TextLink : un projet conjoint de base de données en ligne, qui montre et relie des lexiques, aussi bien existants que créés récemment, de connecteurs discursifs dans plusieurs langues. Nous commençons par considérer la définition et la délimitation de la classe des connecteurs qui devraient être inclus dans une telle ressource, et nous présentons l'information syntaxique, sémantico-pragmatique et lexicographique que nous avons recueillie. D'autre part, l'implémentation technique de cette base de données et les fonctionnalités de recherche qu'elle permet sont aussi décrites. Nous discutons de quelle manière l'intégration multilingue de plusieurs lexiques de connecteurs apporte une valeur ajoutée aux chercheurs en linguistique et aux autres utilisateurs qui s'intéressent aux connecteurs, en permettant de comparer plusieurs langues et de relier directement les connecteurs dans différentes langues. Pour finir, nous donnons des indications quant à une possible extension future en termes d'ampleur (par exemple, en ajoutant des lexiques pour de nouvelles langues) et de profondeur (en augmentant l'information qui est donnée pour chaque connecteur et en renforçant les liens entre lexiques).

Mots clés : *connecteurs discursifs, lexique, ressources multilingues,* linking

We gratefully acknowledge the work of Felix Dombek, who implemented the web app, built many scripts for converting lexicon formats, and also performed various manual mappings and error corrections in some of the source lexicons. Later, Peter Bourgonje took over the job of technical administration of the app, and contributed to the integration of new lexicons. Further, we would like to thank the two anonymous reviewers for their helpful comments. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) [German Research Foundation] – Project number 317633480 – SFB 1287, Project A03.

1. Introduction

1 Among the discourse relational devices (DRDs) that the TextLink network[1] has been concerned with, connectives play a central role in describing the coherence and cohesion of written text. (In spoken language they are used as well, but here, additional types of discourse markers or particles need to be accounted for.) Besides anaphors, connectives are cohesive instruments *par excellence*, as they provide an explicit semantic relation between two – usually adjacent or embedded – spans of text. This connection is commonly called a *coherence relation* or *discourse relation* when looking at it from the perspective of text function, while from the viewpoint of the connective the relation is called its *sense*. In the literature, various proposals for inventories of such relations or senses have been made, such as those by Mann and Thompson (1988), Sanders et al. (1992), Asher and Lascarides (2003), or Prasad et al. (2008).

2 The relationship between connective and relation/sense is by no means simple, however. While there are cases of widespread agreement (*although* signals a Concession), many connectives can be seen as ambiguous (*since* can signal a temporal connection or a causal one) or as vague (*but* can signal Contrast or Concession, but what is the difference between these two, exactly, and can it be distinguished in every context of use?). Further, it is well-known that some connectives can operate on different linguistic "levels" (cf. Sweetser, 1990), as exemplified by a causal connective that relates in fact to a speech act rather than to a proposition: *Are you busy now? Because you should read this great article in* The Times. Besides these issues of semantics and pragmatics, there are also many open questions concerning the exact borderline of the class of "connective": which terms fit in, and which do not, and why? We believe that progress on these matters is more easily made when *inventories*, i.e., lists of connectives with linguistic descriptions, are freely available for as many languages as possible, so that researchers can inspect, consider, possibly disagree with boundaries of inventories or with linguistic descriptions; provide arguments for changes; and contribute to further improving our understanding of what these interesting lexical items actually do, and how that should be documented.

3 When the TextLink network started surveying inventories of discourse-oriented corpora and of DRD lexicons in 2013, only two lexicons were found. The general

[1] See: http://textlink.ii.metu.edu.tr.

consensus was that it would be desirable to have them for more languages, and so members were encouraged to consider contributing a resource for their favorite language. A few years later, in 2017, the web-based interface Connective-Lex[2] went online with lexicons for four languages; the entries in these lexicons gave some fundamental information about orthography, syntactic category, potential ambiguity and coherence relations. It seemed advisable to be modest on the level of detail in order to make it reasonably simple to build a lexicon for a new language from scratch, so that it could be integrated into the system. Today the system has nine different lexicons, and at least one more has been promised for 2019. We thus consider the above-mentioned goal of "providing motivation" as by and large accomplished. In this paper we motivate and present this multilingual lexicon and its design decisions, and also begin to address the follow-up goal of successively enriching the information provided in connective lexicons, so that it becomes a more valuable resource for studying discourse phenomena both within a language and across multiple languages. After all, much more is known about connective behavior in many languages, which needs to be systematized; besides, much more is still to be discovered (some points for future work will be mentioned in Section 4.2).

4 In short, for the near future, Connective-Lex should be expanded not only in terms of breadth (by adding new languages) but also in terms of depth (by providing more, or more exact information). However, the two goals are not trivial to pursue in parallel: any changes on "depth" need to be done carefully in order to preserve – or even enhance – the comparability between languages.

5 While the various lexicons assembled in Connective-Lex have already been individually introduced in earlier publications, the main contribution of the present paper is to present, for the first time, the joint online database project. We explain what the system does, and how it can be used, so that potential users can check whether it is helpful for their research, and prospective developers of new lexicons know what needs to be done to add them to the system. In addition, as mentioned earlier, we take a step back and reflect on what has been done so far, and what should be done next.

6 In Section 2, we first provide background information by reviewing the origin of the first connective lexicons, and by summarizing some earlier research on the crosslinguistic comparison of connectives. Section 3 then introduces Connective-Lex, both from the perspective of a user and that of a potential lexicon developer. In Section 4, we provide a more detailed investigation of the underlying linguistic questions, viz. the definition and demarcation of the class of connectives, and the information provided in lexical entries. Here, we go beyond the status quo and sketch some possible directions for extensions. Section 5 turns to the multilingual perspective and examines possibilities for improving the support for contrastive studies. Finally, in Section 6 we draw some conclusions and provide an outlook.

[2] See: http://www.connective-lex.info.

2. Related research

2.1. Lexicons of connectives

7 The first machine-readable lexicon of discourse connectives was the German DiMLex, whose version 1 was introduced by Stede and Umbach (1998). The main motivation was to use a declarative resource as a component for software that needs knowledge about discourse relations in order to:

- generate text: when a system produces natural language from underlying structured data or a knowledge representation system, and it needs to express a temporal sequence relation between events, it would select from a set of similar connectives. The first application was Techdoc (Rösner & Stede, 1994), which produced maintenance instructions of car manuals. Example: *Park the car on level ground* ***and*** *switch the engine off.* ***Then,*** *check the engine oil*;
- understand text: when a system is supposed to construct a discourse representation, it can exploit connectives as hints on what relations may hold between adjacent units of text. The first applications were the parsers by Hanneforth et al. (2003) and by Reitter (2003), which worked on German text and built trees in the spirit of Rhetorical Structure Theory (RST – Mann & Thompson, 1988).

8 Shortly after presenting the first version of DiMLex, interactions started with the group of linguistics researchers at the Institut für Deutsche Sprache (Mannheim), who were working towards publishing the voluminous *Handbuch der deutschen Konnektoren* (Pasch et al., 2003). As a result of this collaboration, some changes were made to DiMLex, and more connectives were added, resulting in a set of 175; this is documented in Stede (2002). From the outset, DiMLex had been confined to what is nowadays called *primary* connectives, i.e., the inclusion/exclusion decision was made largely in line with the definition by Pasch et al. (2003), which we discuss in detail in Section 4. In contrast, multiword units that are compositional, modifiable and inflectable (e.g., *for this reason*), which today are often called *secondary* connectives, are deliberately not covered (see, e.g., Danlos et al. [2018], or the discussion of *alternative lexicalizations* by Prasad et al. [2008]).

9 The contents of a DiMLex entry in that early version were:

- orthography: list of alternative spellings (if any);
- syntax: category (subordinating conjunction, coordinating conjunction, adverbial, preposition); phrasal status (one or more words; continuous or not); possible positions within the sentence (this can be quite elaborate, especially for German adverbials; we followed the feature-based approach of Pasch et al. [2003]); possible linear orders of the two arguments;
- semantics: coherence relation or sense; modifiability by focus particles; certain idiosyncratic features applying only to certain connectives (presuppositions, style);

- ambiguity information: whether the word also has a non-connective reading;
- examples for the various readings.

10 To support automatic processing, DiMLex was represented in an XML (Extensible Markup Language) format, which we will illustrate in Section 3 (in the slightly different form as used today in Connective-Lex).

11 Also in the early 2000s, a tool for the semi-automatic annotation of connectives was presented (Stede & Heintze, 2004): ConnAnno reads a reduced version of DiMLex, spots candidate words in the input text and highlights them; if the user confirms that one is a connective, the system uses simple surface rules to suggest possible arguments (on the basis of the syntactic type of the connective), which the user can in turn either confirm or correct.

12 More recently, Scheffler and Stede (2016) built a significant extension of DiMLex: 100 new entries were added, and for the new lexicon of 275 connectives, potential senses of the PDTB-3 hierarchy (Penn Discourse Treebank, version 3 – Webber et al., 2019) were determined by means of a corpus analysis (50 random samples per word, taken from the DWDS corpus[3]).

13 DiMLex and its XML format inspired the design of the French LexConn (Roze et al., 2012), for storing the orthography, syntactic category, and sense of French connectives. The first version of the lexicon was obtained by compiling a list of elements belonging to the syntactic categories described above, and this list was manually filtered. This resulted in a lexicon with 325 entries. The sense inventory was modeled on the basis of Segmented Discourse Representation Theory (SDRT – Asher & Lascarides, 2003). Later on, a corpus annotation project (Afantenos et al., 2012) extended the lexicon with an additional 30 entries. The underlying corpus contained about 18,000 sentences, in which roughly 10,000 connective tokens were annotated; on this basis, the authors now consider the lexicon to be complete. In contrast, DiMLex so far has only been applied to a relatively small corpus of 175 German newspaper editorials (Stede & Neumann, 2014) with 1,100 connective tokens. Since that corpus is also genre-specific, the empirical basis of LexConn is on the whole considerably broader.

14 Enabled by a TextLink short-term scientific mission, Colinet (2015) undertook a detailed comparison of DiMLex and LexConn. She pointed out a difference in the criteria used to demarcate the class of connectives: while the criteria used for DiMLex also treat certain nominalized verbs as potential arguments for connectives, LexConn restricts arguments to full clauses. A consequence is that DiMLex accepts more prepositions than LexConn, which includes only prepositions taking infinitival clauses (of which French has many more than German). A second difference is that LexConn has a number of relatively-lexicalized prepositional phrases (for example, *à l'instant où* [at the instant when]), whose German correspondents are compositional phrases which therefore do not pass the criteria of non-inflectability and non-modifiability (see Section 4.1).

[3] DWDS – *Digitales Wöterbuch der deutschen Sprache*: https://www.dwds.de.

2.2. Multilingual/contrastive studies on connectives

15 In linguistics, it seems that the interest in connectives increased considerably in the 1990s, which may be related to the publications on approaches to defining and taxonomizing coherence relations starting in the 1980s. In this section, we point to some studies that took a decidedly multilingual (or contrastive-linguistic) perspective – with much recent work having originated in the TextLink network – and distinguish them according to the primary goals of the work. These goals are of course not disjoint, as many studies may follow multiple lines; here we try to group approaches by their predominant motivation.

2.2.1. Insight into coherence relations

16 For research on coherence relations, it is methodologically not easy to demonstrate the cognitive or linguistic "reality" of such relations. While it is generally accepted that connectives serve as (more or less clear) signals of such relations, some researchers argued that parallel findings on different languages can be taken as stronger evidence for their (language-independent) role in cognition. In this vein, Knott and Sanders (1998) employed the connective substitution test of Knott and Dale (1994) for producing taxonomies of English and Dutch connectives in four semantic areas (positive and negative causality, positive and negative addition). They examined in what way the non-/substitutability in certain contexts can be explained by cognitive-linguistic features and argue that those features that had been used in the Cognitive Approach to Coherence Relations (CCR – Sanders et al., 1992) are appropriate for the task. Recently, working in the same framework, Hoek et al. (2017) studied the translations of English connectives in four languages of the Europarl corpus (Koehn, 2005) in order to determine factors for the decision whether to also use a connective in the target language or to leave the relation implicit. This question was also investigated in a study by Zufferey (2016). There are also various studies that center on one relation specifically and examine variants of its realization in different languages, for instance the corpus-based work by Grote et al. (1997) on concessions in English and German, which led the authors to propose three different subclasses of the Concession relation.

2.2.2. Translation of connectives

17 Quite a lot of work focused explicitly on the crosslinguistic mapping of connectives, without resorting to coherence relations as a possible inter-lingual backbone. Two recent examples are the investigation of translating *however* into Lithuanian by Mazeikiene and Vaiciuniene (2016) and the study by Nedoluzhko and Lapshinova-Koltunski (2018), who used a parallel corpus to determine the distribution of translation equivalents of a range of German pronominal adverbs in English, Czech and Russian. In machine translation (MT), Meyer and Poláková (2013) presented an MT experiment from English to Czech using a gold-standard of manually annotated discourse connectives, and achieved better translations. They illustrate difficult cases of translation, such as *meanwhile*, which is ambiguous between

a temporal and a contrastive sense, and was wrongly translated, in a contrastive context, with a Czech connective that has only a temporal sense.

2.2.3. Semantics and pragmatics of connectives

18 A similar line of research also compared semantically similar connectives across languages, but focused more on gaining insights into the linguistic description of connective meaning, with translation equivalence being only a secondary aspect. One example is Behrens and Fabricius-Hansen (2002), who examined Norwegian, German and English connectives signaling an Elaboration, analyzed their syntactic context, and found, for example, that English *by*-clauses and German *indem*-clauses have different scope properties: *indem*-clauses fall outside the scope of the negator in the matrix clause, while *by*-clauses have narrow scope, i.e., they attach to the propositional nucleus of the matrix clause. Using the same corpus-based research paradigm, Mortier and Degand (2009) focused on French *en fait* [in fact] and Dutch *eigenlijk* [actually], gathered statistics on their collocations, signaled relations, etc., and obtained a detailed "relational semantic field" that contributes to explaining under what circumstances the two connectives have the same or a different meaning. In a similar way, Zufferey and Cartoni (2012) examined English and French causal connectives and found a number of fine-grained features implicated in their conditions of (monolingual) use and their translation.

2.2.4. Generating lists of connectives

19 Finally, during the last 10 years another research direction has become quite prominent: inducing lists of connectives from parallel corpora, where connectives are (manually or automatically) annotated in one language, and then automatically projected to the other one. Versley (2010) trained a connective classifier on the English PDTB, then automatically tagged part of the English Europarl corpus (Koehn, 2005), and projected the connectives to the (automatically-aligned) German section, so that candidates for German connectives were induced. In similar ways, also starting from existing labeled data for English, Zhou and Xue (2012) built a list of Chinese connectives, Hajlaoui and Popescu-Belis (2012) an Arabic one, and Laali and Kosseim (2014) a French one. Recently, Bourgonje et al. (2017) used Europarl to extend the information in the existing German and Italian lexicons (DiMLex, LICO): they found additional connectives that ought to be added to the respective lexicons, and they also studied the mapping between the annotated senses, in order to find areas of overlap in readings, and to compare the degrees of connective ambiguity in the two languages.

3. Design and implementation of Connective-Lex

20 The web database Connective-Lex was designed with two central goals in mind: it should provide a search functionality that allows users to browse related connectives both within and across languages; and it should allow for easily adding new lexicons (i.e., new languages) to the system. The latter goal on the one hand translated into

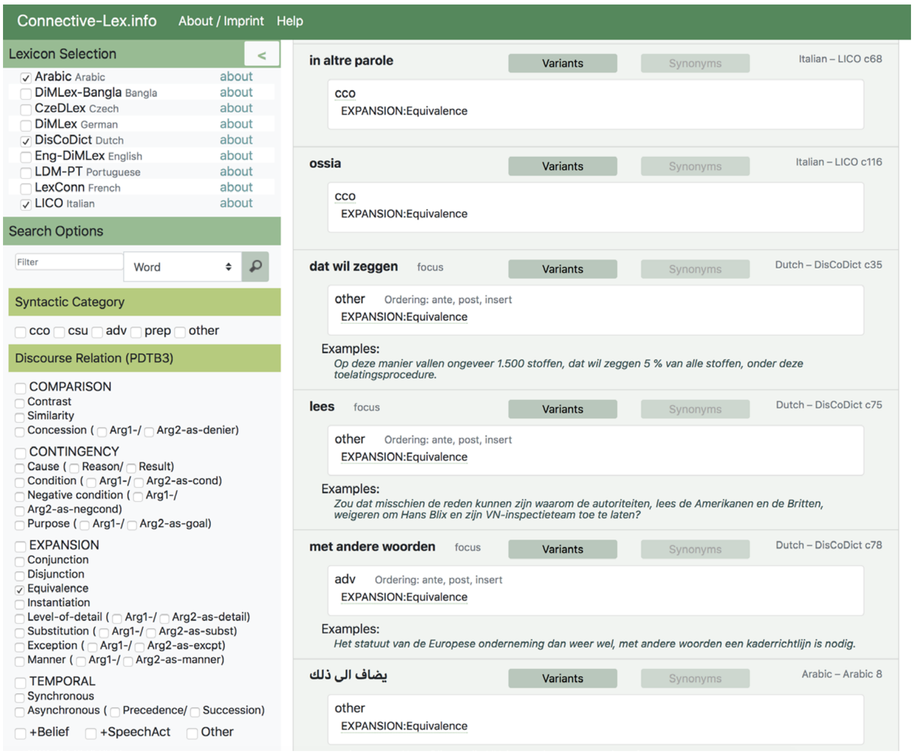
a purely technical requirement on the web app, which should notice the presence of a new lexicon in the resource directory and smoothly integrate it into the search front end. On the other hand, since at this stage the emphasis is on encouraging the development of new lexicons, we kept the contents of entries to a minimum (in order to reduce the effort of building a lexicon for a new language from scratch). At present, lexicons for the following languages are available in Connective-Lex: Arabic (Keskes et al., 2014), Bangla, Czech (CzeDLex; Mírovský et al., 2017), Dutch (DisCoDict; Bourgonje et al., 2018), English (Eng-DiMLex; Das et al., 2018), French (LexConn; Roze et al., 2012), German (DiMLex; Stede, 2002), Italian (LICO; Feltracco et al., 2016), and Portuguese (LDM-PT; Mendes et al., 2018). Pointers to the original source lexicons are provided on the Connective-Lex website (see the "About" links).

3.1. Minimal connective descriptions for the search facility

21 The goals of relative simplicity and of language-neutrality led us to define only a small set of obligatory attributes that need to be known about a connective in order to ensure compatibility with the database. This is, first, the syntactic category, where we use the following inventory: coordinating conjunction, subordinating conjunction, adverbial, preposition (see Section 4 for details). Second, we require that connectives come with one or more senses from the PDTB3 tagset (Webber et al., 2016; Webber et al., 2019). This choice was made because the PDTB corpus is the largest resource available with annotated connectives, and the tagset has therefore been extensively tested on empirical data and in agreement studies. The database uses the full three-level hierarchy of tags, and connectives may be described on any level. The additional attributes ("+Belief" and "+SpeechAct") in PDTB corpus annotation are associated with one of the arguments of the relation; nevertheless, these features are also being used to describe connectives in some lexicons, and therefore we also account for them.

22 Syntax and PDTB3 sense are two of the filter settings of the web app that allow for constraining a search; the other two are the language(s) to be considered and a free text term that is matched against substrings of connectives. For instance, when the language is set to Italian and the string *altri* [other] is entered in the search field, the system returns the entries for the conjunctions *altrimenti* [otherwise] and *in altre parole* [in other words]. The latter results from the fact that LICO provides orthographic variants of that canonical form, one of which is *in altri termini* [in other terms]. Such variants can be displayed if any are available; likewise, there may be synonyms, either in the same or in a different language. For example, since the Italian LICO was modeled closely after the German DiMLex, it also provides pointers to closely related German connectives. (Hence, the term "synonym" is used in a rather wide sense throughout this paper, also covering near-synonyms or "plesionyms", as well as translation correspondents.)

23 A different search scenario checks for connectives in multiple languages that express the same relation. For illustration, Figure 1 shows a screenshot of Connective-Lex with part of the result found for "Equivalence" in Italian, Dutch and Arabic.



Lexicon Selection

☑ Arabic Arabic
☐ DiMLex-Bangla Bangla
☐ CzeDLex Czech
☐ DiMLex German
☑ DisCoDict Dutch
☐ Eng-DiMLex English
☐ LDM-PT Portuguese
☐ LexConn French
☑ LICO Italian

Search Options

Filter | Word

Syntactic Category

☐ cco ☐ csu ☐ adv ☐ prep ☐ other

Discourse Relation (PDTB3)

☐ COMPARISON
☐ Contrast
☐ Similarity
☐ Concession (☐ Arg1-/ ☐ Arg2-as-denier)

☐ CONTINGENCY
☐ Cause (☐ Reason/ ☐ Result)
☐ Condition (☐ Arg1-/ ☐ Arg2-as-cond)
☐ Negative condition (☐ Arg1-/
☐ Arg2-as-negcond)
☐ Purpose (☐ Arg1-/ ☐ Arg2-as-goal)

☐ EXPANSION
☐ Conjunction
☐ Disjunction
☑ Equivalence
☐ Instantiation
☐ Level-of-detail (☐ Arg1-/ ☐ Arg2-as-detail)
☐ Substitution (☐ Arg1-/ ☐ Arg2-as-subst)
☐ Exception (☐ Arg1-/ ☐ Arg2-as-excpt)
☐ Manner (☐ Arg1-/ ☐ Arg2-as-manner)

☐ TEMPORAL
☐ Synchronous
☐ Asynchronous (☐ Precedence/ ☐ Succession)

☐ +Belief ☐ +SpeechAct ☐ Other

in altre parole Italian – LICO c68

cco
EXPANSION:Equivalence

ossia Italian – LICO c116

cco
EXPANSION:Equivalence

dat wil zeggen focus Dutch – DisCoDict c35

other Ordering: ante, post, insert
EXPANSION:Equivalence

Examples:
Op deze manier vallen ongeveer 1.500 stoffen, dat wil zeggen 5 % van alle stoffen, onder deze toelatingsprocedure.

lees focus Dutch – DisCoDict c75

other Ordering: ante, post, insert
EXPANSION:Equivalence

Examples:
Zou dat misschien de reden kunnen zijn waarom de autoriteiten, lees de Amerikanen en de Britten, weigeren om Hans Blix en zijn VN-inspectieteam toe te laten?

met andere woorden focus Dutch – DisCoDict c78

adv Ordering: ante, post, insert
EXPANSION:Equivalence

Examples:
Het statuut van de Europese onderneming dan weer wel, met andere woorden een kaderrichtlijn is nodig.

يضاف الى ذلك Arabic – Arabic 8

other
EXPANSION:Equivalence

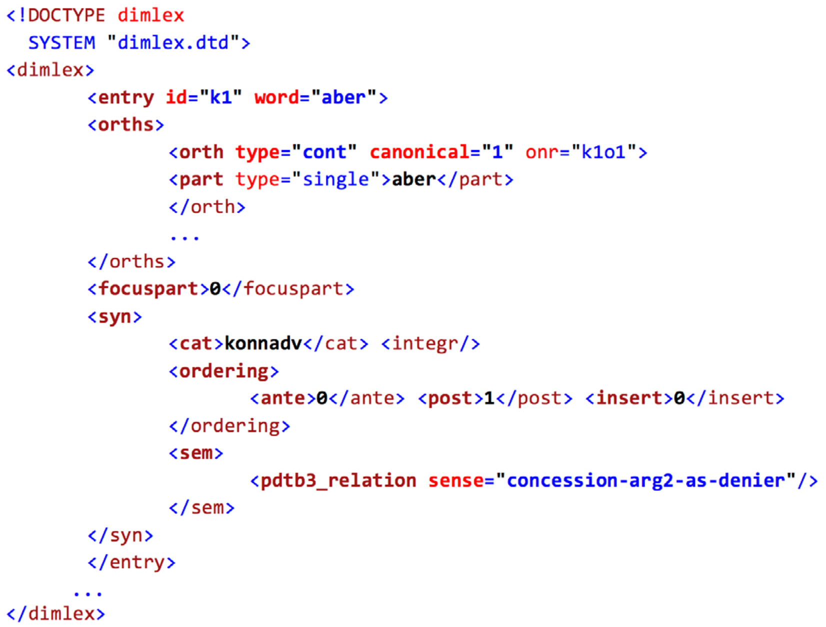
Figure 1 – Screenshot of Connective-Lex

3.2. Integrating lexicons

24 The XML format for importing lexicons is a reduced form of the original DiMLex format; Figure 2 sketches the basic structure of an entry (for brevity, we have omitted several attributes)[4]. After the identifier, the <orths> tag can list different spelling variants[5], one of which should have the value 1 for the "canonical" attribute – this is the spelling shown when the entry is displayed (others appear upon clicking "variants"). Every variant is described for its complexity: single-word or multiword, and continuous (one or more adjacent tokens) or discontinuous (as in *if… then*). Part1 is provided in either case, while Part2 is given only for discontinuous connectives. This structure allows for a single-word connective to have also a multiword synonym sharing the same syntactic and semantic behavior; the two can thus be listed within the same entry.

[4] The DTD (document type definition) for the format is available from the Connective-Lex website (see "About").

[5] In the German DiMLex, these are used to capture different spellings of an "umlaut", or to record variants as they resulted from the German spelling reform in 1996. This would also be the place to add abridged variants used in chat and social media, if desired.

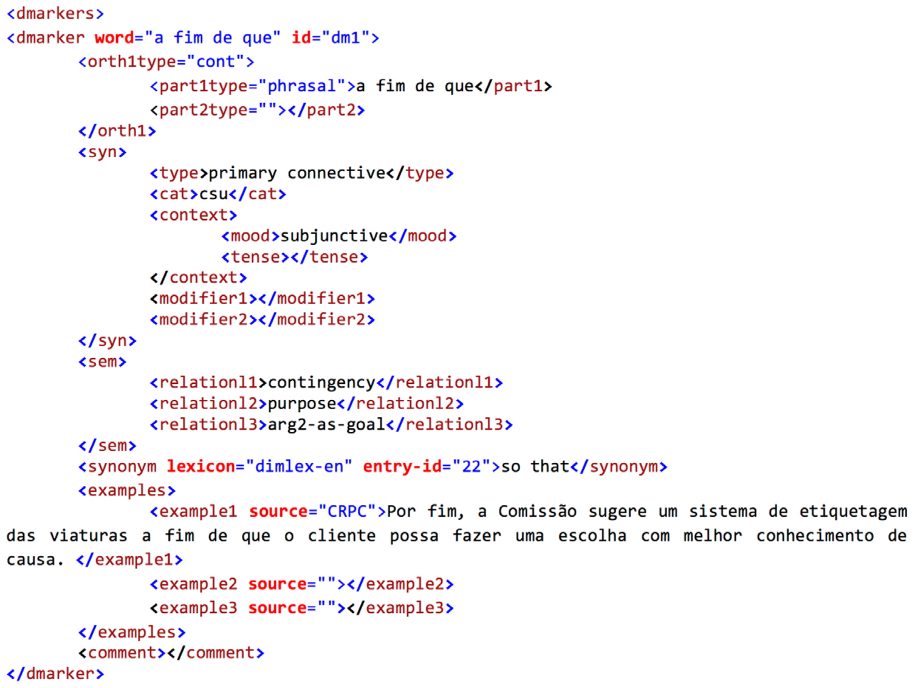


```
<!DOCTYPE dimlex
  SYSTEM "dimlex.dtd">
<dimlex>
        <entry id="k1" word="aber">
        <orths>
                <orth type="cont" canonical="1" onr="k1o1">
                <part type="single">aber</part>
                </orth>
                ...
        </orths>
        <focuspart>0</focuspart>
        <syn>
                <cat>konnadv</cat> <integr/>
                <ordering>
                        <ante>0</ante> <post>1</post> <insert>0</insert>
                </ordering>
                <sem>
                        <pdtb3_relation sense="concession-arg2-as-denier"/>
                </sem>
        </syn>
        </entry>
      ...
</dimlex>
```

Figure 2 – Abridged version of XML structure used in Connective-Lex entries

25 The binary <focuspart> tag specifies whether the connective can be in the scope of a focus particle (*especially because*) or not (? *especially while*). Then, we embed the semantic description within the syntax entries, of which there can be many when a connective has multiple syntactic functions; this decision will be explained in Section 4. Besides the categories mentioned in the previous subsections, the syntax part can specify ordering constraints for the argument: i.e., whether Arg2 can precede Arg1 (<ante>), follow it (<post>), or be embedded in it (<insert>).

26 Obviously, the integration of a new lexicon is particularly easy when it is modeled from scratch in a very similar way (as done, for example, with the Italian LICO). Otherwise, we have mapped existing resources: for the case of Portuguese LDM-PT, the original source formats were an Excel sheet and a structurally-equivalent XML (produced via a Perl script); the latter was then mapped to the DiMLex XML format by means of an XSL (Extensible Stylesheet Language) sheet. Details can be found in Dombek (2017). For illustration, Figure 3 shows an entry from LDM-PT in the original XML format. Note that the creators made a different design decision for the relation between syntax and semantics: both are on the same level of embedding, i.e., one entry always gives one combination of syntax and semantics. The XSL script thus needs to merge different entries into a single one, iterating over entries for identical words and over their syntactic categories. A very similar situation held for the French LexConn, where it was also important to have the example sentences and (within-language) synonyms appear in the right place in our target structure.



```
<dmarkers>
<dmarker word="a fim de que" id="dm1">
        <orth1type="cont">
                <part1type="phrasal">a fim de que</part1>
                <part2type=""></part2>
        </orth1>
        <syn>
                <type>primary connective</type>
                <cat>csu</cat>
                <context>
                        <mood>subjunctive</mood>
                        <tense></tense>
                </context>
                <modifier1></modifier1>
                <modifier2></modifier2>
        </syn>
        <sem>
                <relationl1>contingency</relationl1>
                <relationl2>purpose</relationl2>
                <relationl3>arg2-as-goal</relationl3>
        </sem>
        <synonym lexicon="dimlex-en" entry-id="22">so that</synonym>
        <examples>
                <example1 source="CRPC">Por fim, a Comissão sugere um sistema de etiquetagem
das viaturas a fim de que o cliente possa fazer uma escolha com melhor conhecimento de
causa. </example1>
                <example2 source=""></example2>
                <example3 source=""></example3>
        </examples>
        <comment></comment>
</dmarker>
```

Figure 3 – Sample entry from LDM-PT (Mendes et al., 2018: 4382)

27 Besides this merging, the other central task for the mapping scripts is to convert syntactic categories (or part-of-speech tags) and sense information. The original DiMLex, for example, uses a slightly more elaborate syntax tagset (*konnadv*, *padv*, *konj*, *subj*, *v2emb*, *postp*, *appr*, *appo*, *apci*, *einzel*), which is being reduced to that of Connective-Lex. Relating different semantic/pragmatic relation sets is more difficult, and we discuss this in Section 4. Technically, we used declarative tables for mapping tagsets, so they can be easily re-used and adapted for the integration of a new lexicon.

28 An additional advantage of using XML formats is that by means of XSLT (Extensible Stylesheet Language Transformation) the information can be straightforwardly converted to HTML (Hypertext Markup Language) for viewing in a web browser. The LexConn developers, for instance, provided such a script from early on, so that their lexicon can be viewed online[6].

3.3. Implementation of the web app

29 As explained briefly in Scheffler et al. (2018) and in detail by Dombek (2017), the web app consists of frontend and backend, and is built on the principles of the

[6] See: http://www.linguist.univ-paris-diderot.fr/-croze/D/Lexconn.xml.

so-called *single-page web application* (SPA) technology, which has several advantages: the frontend resides completely in a HTML page, so there are no page reloads, and all the rendering is done on the side of the client (browser). Accordingly, client-server communication is kept to a minimum, and it is done by sending JSON (JavaScript Object Notation) data. When the client loads some data from the PHP (Hypertext Preprocessor) backend, it updates the document by directly editing the HTML structure (DOM – Document Object Model). All these measures serve to speed up the browsing experience of the user.

30 One area of potential improvement is *responsive design*. So far, the system is built for large screens, and enabling a smooth experience with handheld devices has not been our focus. But this step is on the agenda for the near future.

31 Importantly, the system is designed such that adding a new lexicon is very simple: the lexicon itself (in DiMLex XML), a metadata file and, if necessary, mapping tables for syntactic or semantic tags are added to the data directory in the backend, and the app then automatically integrates the new resource and makes it available to the users.

3.4. Creating compatible connective lexicons

32 Connective-Lex can easily integrate lexicons for additional languages if they adhere to the described basic XML format. Such lexicons can be created in a variety of ways: (i) by manual compilation, (ii) by translating and editing an existing lexicon, or (iii) by extracting connectives from an annotated corpus of discourse structure (e.g., in PDTB format).

33 The most straightforward way for lexicon creation for a new language is the manual compilation of an inventory of discourse connectives, which then has to be completed by adding relevant syntactic and semantic properties. This process can be aided semi-automatically by starting from a list of lexical items that fall into the major syntactic classes for connectives (conjunctions, sentence adverbs, adpositions). For example, the creators of the French LexConn (Roze et al., 2012) started with a list of connective candidates based partly on syntactic criteria.

34 Manual lexicon creation is a very time-consuming task. The process can be speeded up by starting from an existing lexicon for another (possibly related) language, and translating the items to obtain a draft lexicon of connectives that includes not only the items but their potential syntactic and semantic properties, as well. The Italian lexicon LICO (Feltracco et al., 2016) enhanced existing lists of Italian discourse connectives by translating the example sentences in the German DiMLex lexicon and keeping those candidate translations that preserve the sense of the German connective.

35 Finally, an initial list of discourse connectives can be obtained automatically from a corpus annotated in PDTB format. Each explicit discourse relation should be extracted with its syntactic type and sense annotation, to yield a list of connectives with all

major properties. The lexicon of connectives attested in the English PDTB corpus was extracted in such a way[7] and was used as the basis for the English connective lexicon included in Connective-Lex (Das et al., 2018). A corpus-extracted lexicon is not necessarily complete and can be extended manually or by translating from other resources; such extensions were also added to the English PDTB lexicon, thus enlarging it by 50%.

4. Taking stock: toward a multilingual connective database

36 As presented in the previous sections, Connective-Lex provides a multilingual interface to connective lexicons in several languages, but was built on the model of the German DiMLex lexicon. In this section, we will assess the effects of this design and evaluate the fundamental limits of such a multilingual resource. In addition, we compare our "machine lexicon" with human-created resources and identify potential information types that can be added systematically to the existing platform.

4.1. Connective: definition and delimitation

37 The first important decision in the creation of a specific lexicon is which items to include in it. For a lexicon of connectives or discourse markers, it is essential to define exactly what is meant by these terms, since individual researchers have proposed different ways of limiting the space of the phenomena under discussion. Definitions also typically follow language-specific considerations (or are focussed on a small set of languages, as discussed in Section 2). As the starting point for the multilingual interconnected lexicons described here is the German DiMLex, we largely follow their definition of "discourse connectives", based on the work by Pasch et al. (2003). Our definition has four main criteria:

[1] Definition of connective
A lexical item or phrase x is a connective if and only if:
1. x cannot be inflected or modified;
2. the meaning of x is a two-place relation;
3. the arguments of this relation are abstract objects (propositions, events, states, or processes);
4. these arguments are typically expressed as clauses or sentences (but they can also be expressed as groups of sentences, or as noun phrases that denote abstract entities).

38 In addition to single words which have these properties, fixed phrases can be included as connectives if they cannot be modified (see condition 1) and if their internal semantics is thus not compositional. The intention of this condition is to exclude the potentially very large number of freely modifiable phrases that also

[7] It is available at: https://github.com/TScheffler/Connectives.

denote functions over abstract entities from a finite and focussed resource such as a connective lexicon. Freely modifiable phrases include for example many verbs that have denotations which also occur as discourse relations, such as *cause*, *lead to*, *contrast with*, etc. These verbs contribute their relations to the regular sentence semantics. They contribute to the discourse structure of a text (Danlos, 2006), but should not be included in a connective lexicon.

39 This definition closely resembles the definitions for "discourse connectives" assumed in many other individual language lexicons, and it is also compatible with that given by Danlos et al. (2018) for "primary connectives". However, here we ask whether a language-independent definition is possible and well-founded. The English PDTB (Prasad et al., 2008) identifies connectives as items that express binary semantic predicates whose two arguments are abstract objects (Asher, 1993). The connective signals a semantic or pragmatic relation between the two arguments. The PDTB also places strict syntactic restrictions on the items that can be classified as connectives. These syntactic restrictions are of two types. On the one hand, the connective item itself must come from one of only three parts of speech: it must be a subordinating or coordinating conjunction, or an adverbial [8]. On the other hand, the syntactic form of the connective's arguments is also restricted: connective's arguments are assumed to be expressible as clauses. This allows for arguments that are (one or more) sentences, clauses, or verb phrases. Gerund clauses in English also fall under this definition. But any item that can only take noun phrase arguments (even if they express abstract objects such as events) is not considered a connective under this definition.

40 The core semantic and pragmatic properties of connectives carry over crosslinguistically. All connective lexicons in different languages include the criteria 2 and 3 listed above: a connective must express a two-place semantic or pragmatic relation, whose arguments are abstract objects. These criteria are essential to distinguish connectives from many other discourse markers (which may denote one-place relations, such as most sentence adverbs), and to distinguish discourse connectives from other two-place semantic predicates (such as verbs, etc.). These two main criteria are for example included in the lexicons for Arabic (Al-Saif & Markert, 2010 [9]) and Chinese (Zhou & Xue, 2015), which are based on the PDTB definition, but also in the Portuguese LDM-PT (Mendes et al., 2018), the French LexConn (Roze et al., 2012), and the Czech CzeDLex (Mírovský et al., 2017), which start from independent corpus efforts.

41 Specific syntactic restrictions, on the other hand, lead to problems and grey areas when applied simultaneously to many different languages. Fine-grained part-of-speech categories often do not have close matches in another language. Thus,

[8] Connectives that are coordinated with other connectives (*if and when*) or that are modified by a limited range of focus particles (*even if*) are also included.

[9] Note that this is not the Arabic lexicon included in Connective-Lex.

the decisions on what to include in a given corpus annotation or connective lexicon have often been made based on language-specific criteria, such as the restriction to adverbials and conjunctions, and the ability to take clausal syntactic arguments, in the PDTB. In CzeDLex, prepositions are excluded as primary connectives, because connective's arguments are required to have a verbal core. In contrast, many connective lexicons in other languages allow pre- or postpositions, since these items otherwise fulfil the semantic requirements of a connective. An adposition may have a nominalized internal argument, which can often express an abstract object. For example, the Arabic discourse treebank includes prepositions that take nominalized arguments (in the *Al-Mazdar* form), as well as clitics; the French LexConn explicitly includes prepositions as candidates; the Chinese lexicon includes a type of postposition called *localizers*; and the German DiMLex allows for pre- or postpositions with noun phrase arguments that denote abstract objects. In many cases, these types of connectives would correspond to items in English that take gerund arguments and thus may be connectives according to the PDTB definition as well.

42 Completely removing any syntactic constraints from the definition of "connective" will also lead to conceptual problems, however. For example, there are certain types of verbs that can take clauses as arguments. Consider the English verb *depend on*, as in the following sentence:

[2] Whether we will go skiing *depends on* what the weather is like.

43 The verb denotes a two-place semantic relation between two abstract objects, which are expressed as clauses. Still, one would not like to include such a verb as a discourse connective, because it does not relate independent clauses and its meaning is part of sentence semantics and not discourse. For this reason, we add condition 1 to our definition of connectives. It captures the crosslinguistic generalization that connectives typically come from closed word classes that encompass function words. Open class items and phrases such as nouns or noun phrases, and verbs, are to be excluded. These open class items are inflected in many languages and generally allow for free modification. Several individual language connective lexicons try to capture facets of this constraint by excluding items that are partly substitutable, variable, or modifiable (e.g., French LexConn). CzeDLex and LDM-PT distinguish between non-modifiable, fully grammaticalized "primary" connectives, and freer secondary connectives. Only the primary ones match our definition.

44 The chosen definition of the notion "discourse connective" is thus language-neutral as far as possible. It references mainly semantic/pragmatic notions that are independent of language-specific morphosyntactic instantiations (conditions 2 and 3). Condition 4 is not strictly a constraint since it merely describes typical shapes of arguments (without properly restricting them). The only morphosyntactic constraint (condition 1) is deliberately chosen to be broadly applicable. It will lead to specific ranges of categories of items included as discourse connectives based on the properties of each language, but it guarantees a common core and as much overlap as possible.

4.2. Information on lexicon entries

45 As detailed in Section 3, Connective-Lex currently indexes only limited information for the connectives in its lexicons: mainly orthography, basic syntactic information such as part of speech, and the semantic relation expressed. Some manually compiled lexicons are much more comprehensive. Here, we briefly discuss the additional types of information included in linguistic connective lexicons and their possible inclusion in an extension of our machine lexicons.

4.2.1. Syntax

46 Connectives are grouped by their major syntactic category on Connective-Lex. This enables researchers to search for, e.g., coordinating conjunctions in several languages at once. In addition, this grouping is justified by the observation that connectives mainly hail from very few syntactic classes crosslinguistically. Most notably, this includes coordinating and subordinating conjunctions, adverbials, and adpositions. Except for adpositions, which are sometimes excluded for syntactic reasons, these categories of words are present in all existing connective lexicons. Connective-Lex is agnostic with respect to the actual part-of-speech tags assigned to these word categories within each lexicon. Instead, the language-specific tags are mapped onto general syntactic categories in the web-app interface. This is generally unproblematic, except for the English lexicon, whose syntactic information is based on the Penn Treebank part-of-speech tagset, which assigns both subordinating conjunctions and prepositions the same tag IN[10]. Any part-of-speech annotation that is compatible with the universal part-of-speech tagset provided with Universal Dependencies[11] is easily mappable onto the tagset used by Connective-Lex.

47 In addition, many connective lexicons include certain fixed phrases which can be used as connectives. We can distinguish three main categories of such phrasal connectives:

1. phrasal connectives that are syntactically equivalent to single-word connectives, such as adverbial connective phrases: *on the other hand* (English), *abgesehen davon* ([despite that/other than that], German);
2. paired connectives;
3. other grammaticalized phrases.

48 Many phrasal connectives can easily be mapped onto one of the four major syntactic categories included in Connective-Lex, because they play the same syntactic role as single-word connectives. Thus, English *on the other hand* often matches *also*, and German *abgesehen davon* [despite that] behaves like *nevertheless*. They are both listed as adverbials in the multilingual lexicon. Sometimes, complex phrasal

[10] In this case, both syntactic categories adposition and subordinating conjunction are listed as possible, since they cannot be easily distinguished and are often both valid options (Das et al., 2018).

[11] See: http://universaldependencies.org/u/pos/.

connectives also behave like other syntactic types, for example the German *abgesehen davon, dass* (also "despite the fact that", but including the complementizer), which has the syntactic role of a subordinating conjunction.

49 A special phenomenon present in many languages is the case of paired connectives. These connectives always or typically occur paired, one connective introducing each of the two arguments of the relation expressed by the connective pair. Many lexicons in Connective-Lex feature these kinds of connectives, for example English (*either… or*, *if… then*) and German (*sowohl… als auch* [both... and]). In Chinese, paired connectives dominate the range of possible explicit connectives. Since the paired connectives typically exhibit the same syntactic category for both parts, we categorize them the same way we would their individual parts. That is, *sowohl… als auch* is a coordinating conjunction, because each part behaves like one.

50 Finally, each language exhibits idiosyncratic grammaticalized expressions that function as discourse connectives. These expressions frequently consist of more than one word. Sometimes, the two words do not even form a phrase/constituent in the syntactic sense, as with the German connective *sei es* [be it = if], where *es* [it] is arguably the syntactic subject. Since such cases occur on an individual case by case basis in each language, without allowing for crosslinguistic generalizations, we provide the syntactic category "other" for these kinds of (single-word or phrasal) connectives.

4.2.2. Semantics/pragmatics

51 The other central type of information included in the lexicons on Connective-Lex is the possible discourse relations that can be signaled by each connective. Since one discourse connective can have different syntactic instantiations, we subordinate the semantic readings under each syntactic instantiation separately. This is justified by the fact that many discourse connectives denote different semantic relations when they are different parts of speech. Each syntactic option for each connective therefore lists the set of readings available for this connective-part-of-speech combination (see Figure 1).

52 Discourse relations are semantic/pragmatic entities and applicable crosslinguistically. Still, there are many options for inventories of discourse relations ranging from Knott (1996), through the relation taxonomies developed in the RST and SDRT frameworks, to the different variants of PDTB relations (see Section 2.1). Since Connective-Lex assumes a lexical approach to the representation of discourse relations, we chose the PDTB3 relation hierarchy (Webber et al., 2016), which is an improved variant of the well-attested PDTB2 hierarchy. The PDTB3 relations have been used in an annotation project for a diverse range of languages, the TED-Multilingual Discourse Bank corpus (TED-MDB – Zeyrek et al., 2018).

53 In order to provide a crosslinguistically stable search interface for discourse relations, all semantic annotations of existing lexicons must be mapped onto the PDTB3 relation inventory for the purposes of displaying compatible information. We

have developed a unidirectional mapping from PDTB2 relations to PDTB3 relations that are available for the same connectives. For a connective that has been annotated with PDTB2 senses, this mapping lists, as far as automatically possible, the PDTB3 relations that are minimally available for this connective[12]. The mapping table we use is provided in Table 1 in the Appendix. In the spirit of lexicography, which must list all possible options for a given item, PDTB2 relations can also be mapped to several matching PDTB3 relations. Since the French LexConn uses SDRT relations, another mapping links those to PDTB3 (see Table 2 in the Appendix).

54 In mapping the relation schemas, some relations lack corresponding PDTB3 senses. In addition, some connectives may have lacked a clear sense assignment already in the original monolingual connective lexicon. These cases are assigned the relation sense "other" in order to make them searchable on Connective-Lex. For example, the German connective *so oder so* [in any case/anyway] denotes an unconditional relation that is not represented in the PDTB3 sense inventory. Another example is *au moins* [at least] in French.

4.2.3. Lexicographic information

55 Within a given linguistic system, the lexical items, including connectives, may specialize in many specific ways. Some manually compiled lexicons detail this type of valuable information. The Spanish *Diccionario de partículas discursivas del español* (DPDE – Briz et al., 2008)[13] includes detailed, unstructured information on its 210 discourse particles (not including conjunctions and prepositions), including a definition, examples, information on prosody and punctuation, sentence position of the marker, register, variants, translations into other languages, and others. The comprehensive *Handbuch der Konnektoren* (Pasch et al., 2003; Breindl et al., 2014) likewise provides item-level descriptions that include possible sentence positions, meaning and register information, available modifiers or focus particles, and idiosyncrasies for each connective. Some of this information is by necessity language specific. For example, sentence positions for the German connectives are specified according to the topological model of German syntax in Pasch et al. (2003). This model cannot be meaningfully applied to other languages. While specific information on the positioning options of different adverbial connectives in German may be interesting to a researcher, this information cannot easily be integrated into a multilingual resource. In Connective-Lex we have therefore opted to include only core syntactic and semantic information in the item entries and in particular in the search interface. Additional details can be displayed, if available, on a language-specific basis.

[12] Note that this does not necessarily mean that a specific *instance* of a connective annotated with a PDTB2 relation expresses the corresponding PDTB3 relation, but only that a connective which has a certain PDTB2 sense also has the corresponding PDTB3 sense.

[13] The dictionary is available online at: http://www.dpde.es/.

56 Details on register restrictions and possible modifiers or focus particles associated with connectives could however be included in a multilingual lexicon, as well. These specific details are of great interest to researchers investigating connectives, as well as other users of a multilingual resource. Adding normalized genre and register information would add great value to the dictionary over general dictionaries. For example, the popular online dictionary LEO[14] provides the German connectives *da*, *weil* and *denn* as translations of the English causal connective *because* (along with other items, mostly longer phrases). Corpus research has shown that these connectives are not used interchangeably but instead specialize to the level of causality, and also to the mode (spoken or written) or the medium/register (e.g., in German Twitter, *weil* is used almost exclusively; Scheffler, 2014). If available, such register restrictions should be added to a future version of the lexicon. This requires a crosslinguistically applicable set of features for genre, register, and for different types of modifiers. Such inventories should be borrowed from lexicography in future extensions of Connective-Lex.

5. Linking the languages

57 Individual lexicons of connectives are useful for discourse studies and NLP (Natural Language Processing) tasks related to the automatic identification of connectives (Stede & Heintze, 2004; Mendes & Río, 2018) and discourse processing (Lin et al., 2014; Stede, 2014), but the possibility of linking these resources provides additional advantages for research areas such as machine translation (Meyer et al., 2011; Meyer & Poláková, 2013), to develop statistical translation models that are able to operate above the sentence and/or phrase level and that correctly model discourse connectives and coherence relations. Linked resources also provide a very helpful support for manual translation as well as language learning and teaching in CALL (Computer-Assisted Language Learning) systems (Meurers & Dickinson, 2017).

58 As shown in Section 3.1, Connective-Lex enables the users to retrieve, in multiple languages, connectives that express the same sense, or connectives of a specific category. For instance, using the search options “sense:Purpose (Arg2-as-goal)” and “category:csu” (subordinating conjunction) over the German and the Portuguese lexicons retrieves four Portuguese connectives (*a fim de que*, *de forma que*, *de modo a que*, *para que*) and two German connectives (*bis dass* and *damit*). Of these, the first is ambiguous between a temporal and purpose interpretation, and the second has only a purpose interpretation. The search options are therefore quite useful for a contrastive approach. They are nevertheless limited. One such limitation is the result of different inventories of discourse relations, as mentioned in 4.2.2. For instance, the sense “CONTINGENCY:Purpose” is new in the PDTB3 hierarchy, while only “CONTINGENCY:Result” is found in PDTB2. When mapping from PDTB2 to PDTB3, the new relations, such as “Purpose”, could not be automatically mapped.

[14] See: https://dict.leo.org/german-english/because.

As a result, a search in Connective-Lex using "CONTINGENCY:Purpose:Arg2-as-goal" as a search option will retrieve six connectives in LDM-PT and nine in DiMLex (both using PDTB3), but none in Eng-DiMLex (see Appendix for the mapping table and Section 4.2.2 for discussion; in future versions of Connective-Lex the goal is to overcome such differences in the sense hierarchy). Also, when connectives have been automatically extracted from annotated corpora, the listing of senses might need revision, and this will have negative effects when searching multilingual data.

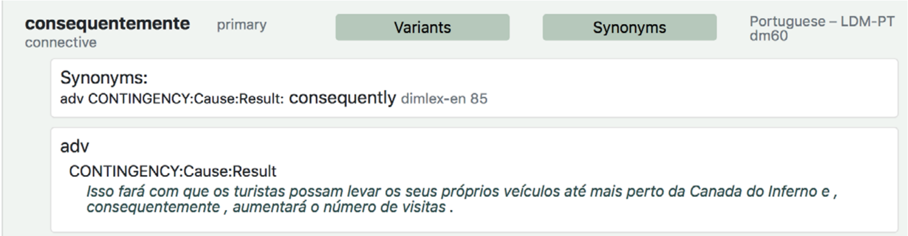
59 Some lexicons in Connective-Lex are further linked through "synonym" pointers, provided during the manual creation process of those lexicons. For instance, the connective *a fim de que* is labeled in Portuguese as a subordinating conjunction ("csu") with the PDTB3 sense "CONTINGENCY:Purpose:Arg2-as-goal". It is linked to the entry *so that* in Eng-DiMLex, labeled as "CONTINGENCY:Result:Arg2-as-result". Although the senses of each connective differ, the linking establishes that they are near-synonyms.

60 One important future step is to systematize this linking between the lexicons in Connective-Lex and to enable a "top-down" view on the similarities and differences between the languages. This task raises the issue of the way in which the linking of connectives should be put into place. Two possible approaches are: (i) to link all pairs of languages or (ii) to use English as the pivot language. Linking the lexicons through English as a pivot language is possibly the best solution in Connective-Lex, considering that the set of lexicons is quickly enlarging and that it would prove difficult to link all pairs of languages. In Connective-Lex, the Portuguese lexicon includes a synonym linking to the English lexicon, while the Italian links to a synonym in the German lexicon (the Italian lexicon was translated from the German DiMLex, cf. Section 3.3). In the following subsections, we will explore the linking of the Portuguese and English connectives, as a source of information for the general goal of using English as the pivot language in Connective-Lex.

5.1. Linking connectives in the DiMLex structure

61 Considering the set of features and the internal structure of the connective's entry, what information should be linked? The word form of the connective is clearly not enough because most connectives are ambiguous between one or more rhetorical senses. For instance, in 2.2.2, we refer to the ambiguity of *meanwhile* and the difficulty in translating to Czech (Meyer & Poláková, 2013). It is thus not enough to map the word forms in two languages, but rather it is necessary to map both the word form and its sense (or one of its senses).

62 As mentioned in Section 3.2, in LDM-PT, each entry corresponds to an association of word form/category/sense. So, the same word form will occur in two different entries if it has two different categories or two different meanings, similarly to the structure of LexConn. The Portuguese lexicon in Connective-Lex includes one or more English synonyms for each word form/sense pair. We choose,



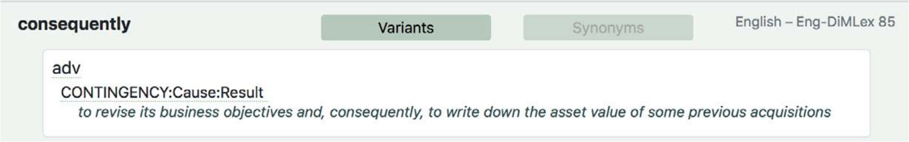
consequentemente primary
connective

Portuguese – LDM-PT
dm60

Synonyms:
adv CONTINGENCY:Cause:Result: consequently dimlex-en 85

adv
CONTINGENCY:Cause:Result
Isso fará com que os turistas possam levar os seus próprios veículos até mais perto da Canada do Inferno e , consequentemente , aumentará o número de visitas .

Figure 4 – Entry for the connective *consequentemente* in LDM-PT



consequently

English – Eng-DiMLex 85

adv
CONTINGENCY:Cause:Result
to revise its business objectives and, consequently, to write down the asset value of some previous acquisitions

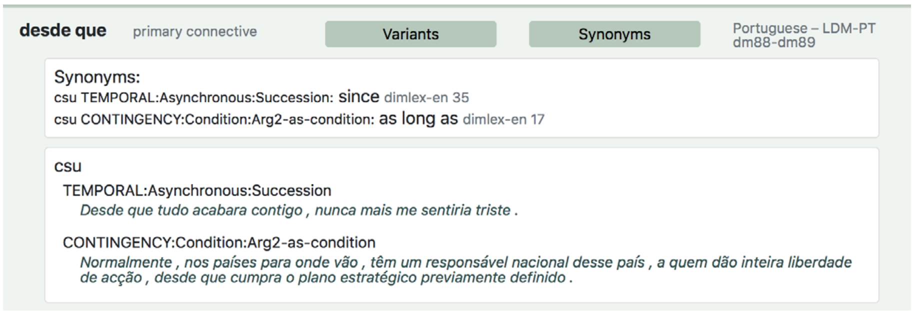
Figure 5 – Entry for the connective *consequently* in Eng-DiMLex

when applicable, one of the entries of the Eng-DiMLex, compiled from data from the PDTB, and provide the unique identifier of the connective in the English lexicon (Das et al., 2018).The English synonym has an equivalent sense but should also occur in the same context, so it should be the same category when applicable. As the near-synonyms are attached to a specific sense, they are inserted into the DiMLex structure as children of the attribute “sem”. The synonym relation then connects two specific word forms/categories/senses.

63 When mapping LDM-PT to the DiMLex format, the individual entries for ambiguous connectives were merged by grouping them, first by word form, then by word class (see Section 3.2). As a result, the synonyms are now presented under the main entry of the Portuguese connective, but they are still linked to a specific category and sense of the connective in LDM-PT. For instance, the adverb *consequentemente* with the sense “CONTINGENCY:Cause:Result” is linked to the English adverb *consequently* in the Portuguese lexicon (see Figure 4). The entry in the English lexicon (see Figure 5) shows that the connective has the same category and the same sense as the Portuguese one. Additionally, both examples in the two languages illustrate the use of the connective in a similar position, introducing the proposition interpreted as result.

5.2. Linking ambiguous connectives

64 The linking is a bit more complex when the connectives have different senses. In these cases, each sense of the Portuguese connective is linked to a specific English synonym. This is illustrated by *desde que*, which is linked to *since* and *as long as* (see Figure 6). The connective is linked to the English entry and not to a specific sense of the English connective. In most cases, the same sense is listed in the Portuguese



desde que primary connective

Portuguese – LDM-PT
dm88-dm89

Synonyms:
csu TEMPORAL:Asynchronous:Succession: since dimlex-en 35
csu CONTINGENCY:Condition:Arg2-as-condition: as long as dimlex-en 17

csu
TEMPORAL:Asynchronous:Succession
Desde que tudo acabara contigo , nunca mais me sentiria triste .

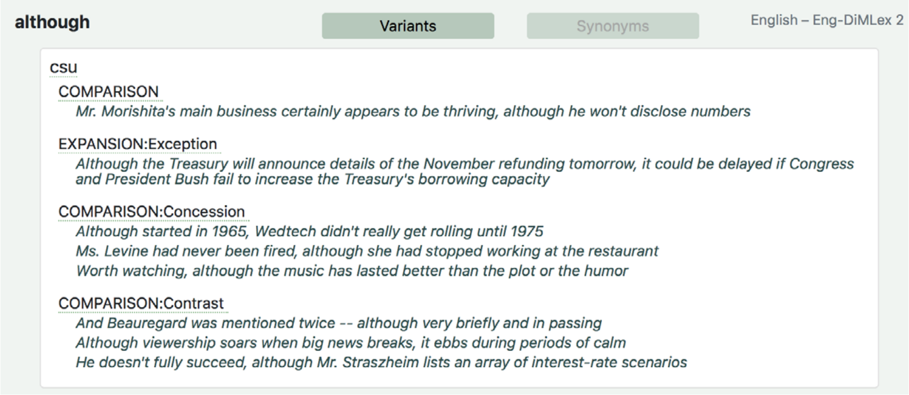
CONTINGENCY:Condition:Arg2-as-condition
Normalmente , nos países para onde vão , têm um responsável nacional desse país , a quem dão inteira liberdade de acção , desde que cumpra o plano estratégico previamente definido .

Figure 6 – Entry for the connective *desde que* in LDM-PT

and English entries, so the linking with the proper sense is straightforward. However, in the future we should consider the possibility of linking to a specific sense of the English connective, rather than to the full entry. This would guarantee that the linking is fully bidirectional, as the information on word form and sense could be retrieved in both directions.

65 An important aspect in the linking process is the underlying perspective towards ambiguity that is adopted by each lexicon. The interpretation of the connectives in context may vary, as a result of the linguistic co-text. On the one hand, a lexicon fully derived from corpus annotation will tend to present a larger set of senses per connective, as a reflection of the decisions made by the annotator to capture the interpretation of the sentence (see Section 3.3 for different methods of creating lexicons of connectives). On the other hand, a lexicon based on corpus data and reference grammars and dictionaries will involve the difficult task of deciding when to include an additional sense. The Portuguese lexicon was derived from the TED-MDB corpus but was then revised to avoid listing a high number of senses in the entries, and the perspective was conservative. Therefore, many entries that are labeled with a single sense in LDM-PT will link to a connective that is highly ambiguous in the English lexicon. For instance, the Portuguese subordinating conjunction *ainda que* is labeled with the single sense “Arg2-as-denier” and is linked to the subordinating conjunction *although*, which has four different values in the lexicon (see Figure 7). Notice, however, that the high number of senses reflects the difficulty in deciding between “Contrast” and “Concession”, which leads, in some contexts, to the choice of the higher-level sense “COMPARISON”.

66 Another example is the highly ambiguous coordinate conjunction *e* [and]. It is labeled with three senses in Portuguese (“Conjunction”, “Result”, “Precedence”), while the English equivalent connective *and* is listed with the same three senses, plus five additional ones (“Condition”, “Contrast”, “Instantiation”, “Concession”, “Arg1-as-detail”) and two higher level senses (“EXPANSION” and “level-of-detail”). Consequently, the Portuguese entry only links to a few of the senses of the connective *and*.



although English – Eng-DiMLex 2

csu

COMPARISON

Mr. Morishita's main business certainly appears to be thriving, although he won't disclose numbers

EXPANSION:Exception

Although the Treasury will announce details of the November refunding tomorrow, it could be delayed if Congress and President Bush fail to increase the Treasury's borrowing capacity

COMPARISON:Concession

Although started in 1965, Wedtech didn't really get rolling until 1975
Ms. Levine had never been fired, although she had stopped working at the restaurant
Worth watching, although the music has lasted better than the plot or the humor

COMPARISON:Contrast

And Beauregard was mentioned twice -- although very briefly and in passing
Although viewership soars when big news breaks, it ebbs during periods of calm
He doesn't fully succeed, although Mr. Straszheim lists an array of interest-rate scenarios

Figure 7 – Partial entry for the connective *although* as “csu” in Eng-DiMLex

5.3. Linking using the *category* attribute

67 The examples above have shown that the word form is not enough to ensure the linking between connectives and that it had to be paired with the sense. One question is whether there are any other constraints to be taken into consideration when linking connectives.

68 Linking does not strictly involve the association of connectives of the same category. However, these categories reflect certain syntactic properties that may be important to keep in the two languages. For instance, a subordinating conjunction will be followed by a finite clause, while prepositions introduce nominal phrases (and non-finite clauses in Portuguese) and adverbs are usually syntactically detached and may occur in different positions. To ensure that the connective is used in the same context, keeping the same category is the preferred option. The importance of the attribute *category* in the lexicons was already pointed out when discussing the retrieval of equivalents in Portuguese and German, two lexicons that are not explicitly linked in Connective-Lex. It makes it possible to retrieve sets of equivalent connectives in terms of sense and category. When explicitly linking two near-equivalent connectives, the goal is to select lexical items from the same category, when applicable. For instance, the subordinating conjunction *ainda que*, labeled as “COMPARISON:Concession:Arg2-as-denier”, is linked to the subordinating conjunction *although*, labeled as “Concession”. This ensures that they both occur in the same set of contexts and in the same positions: they introduce Arg2, they cannot occur inside Arg2, and Arg2 can occur after (Example [3]) or before Arg1 (Example [4]).

[3] a. Este fármaco é usado há já alguns anos de uma forma rotineira nos hospitais portugueses na indução do trabalho de parto e nalguns abortos ao abrigo da lei, *ainda que* esta indicação não esteja prevista na respectiva bula.
(LDM-PT)

b. Ms. Levine had never been fired, *although* she had stopped working at the restaurant.
(Eng-DiMLex)

[4] a. *Ainda que* eu não seja ninguém, não valho menos do que esses tais grandes apóstolos.
b. *Although* started in 1965, Wedtech didn't really get rolling until 1975.
(Eng-DiMLex)

69 The goal was then to select a connective with an equivalent category in Portuguese and in English, but this was not always possible. For instance, *de tal maneira que* [literally "in a such a way that"] is linked to *as a result*. While the Portuguese connective is a subordinating conjunction, highly integrated in the syntactic structure, *as a result* is an adverb and is used in a parenthetical position. The lexicon provides the possible alternative *so*, listed as "csu", among other categories. However, the best match would be *so that*, labeled with the category "other". In this case, keeping the category would not provide the best equivalent connective.

70 Another mismatch is related to the category "preposition". Prepositions introduce infinitive clauses with many different senses in Portuguese, while the set of senses of infinitive clauses seems more restricted in English. Purpose infinitive clauses (introduced by a preposition such as *para* [to], *a fim de* [in order to]) are easily linked to near-equivalent prepositions in Eng-DiMLex. However, infinitive clauses introduced by a preposition may also convey the sense "Reason" in Portuguese, as illustrated in [5] – the connective is the multiword preposition *devido a*. There is no available preposition in English in this context, and the translation of [5] uses instead the conjunction *because* followed by a finite clause. This is another case where the two languages use a different category to express the same sense (and, as a result, a different type of clause in Arg2).

[5] E, no entanto, o ciclo das fases da Lua demonstrava que esta, tal como a Terra, não produzia qualquer luz e apenas brilhava *devido a* reflectir a luz solar.
(LDM-PT)
'And, nevertheless, the cycle of the Moon phases showed that, just like the Earth, it didn't produce any light and only shone because it reflected the sun's light.' (literally: 'because of reflect_inf the sun's light')

5.4. Syntactic, semantic and pragmatic properties

71 Other properties may also be important for the linking process. For example, two synonyms with the same sense and category can differ in terms of the positions in which they may occur in the sentence. For instance, the adverb *mais* is linked to the adverb *moreover* and both are labeled with the sense "EXPANSION:Conjunction". They are also both parenthetical elements, separated by a pause (orthographically

marked by a comma). However, while *mais* only occurs as position-initial, *moreover* may occur in initial position and also internal to Arg2.

72 Other restrictions may apply to synonyms. For instance, the two words *ainda que* and *although* share the same category, sense and position in the sentence, but they differ in terms of mood restrictions. While the Portuguese conjunction requires the use of the subjunctive mood in Arg2, there is no such restriction holding on the Arg2 of *although*. This restriction on mood is not visible in Connective-Lex but the attribute is listed in the LDM-PT lexicon. Restrictions on mood are extremely frequent with Portuguese subordinating conjunctions and would have to be dealt with in automatic processing systems, for instance for machine translation.

73 Finally, pragmatic properties are also to be accounted for when linking connectives. Section 4.2.1 provided examples of connectives specialized in terms of mode (spoken and written) or medium. Indeed, some connectives are preferentially used in formal or informal registers. For instance, the two connectives *pois* and *porque* have the sense "CONTINGENCY:Cause:Reason" but differ because the first one is typically used in formal, written registers. *Pois* also differs because it expresses a justification rather than a cause, but in informal contexts speakers would still rather use *porque* [because] for justification. The same semantic distinction holds between *since* and *because*, and this led us to link *pois* to *since*, and *porque* to *because* (although the difference of register in English would have to be confirmed). Similarly, the prepositions *devido a*, *devido ao facto de*, *em virtude de* [because/due to] all convey a "Reason" relation, but *em virtude de* is clearly used in more formal registers.

5.5. Summary: considerations on linking

74 To conclude, although the search options in Connective-Lex can already provide important insights for a multilingual and contrastive approach, connectives have specific properties that make linking the lexicons more efficient. We discussed the method for such linking and, considering the high number of lexicons in Connective-Lex, propose to use English as the pivot language.

75 As most connectives are ambiguous between different senses, linking merely on the basis of the word form is not enough. The linking of the lexicons has thus to apply to a word form/sense pair. As the lexicons do not all share the same sense inventory, the best solution would be to link directly the sense of the connective in one language to a specific sense of the English connective. In terms of the XML structure of DiMLex this would mean that the linking is established at the level of the attribute "sem". This was not applied in the linking of LDM-PT and Eng-DiMLex: the word form/sense pair in the Portuguese lexicon links to the top-level entry of the English connective, with no specific indication of the sense that it links to. This is not an issue when both lexicons use the same hierarchy of senses, nor if there is in the future a mapping solution between different sense inventories.

76 Another type of information that proves to be important for the linking process is the category of the connective. Since the category restricts many properties of the connective, the linking should also try to keep the same category to guarantee that the connectives do occur in the same syntactic contexts. The examples above show nevertheless that languages differ in terms of the syntactic structures that are used to express some senses and that the category of the synonym in English might differ. While word form and sense are required elements for the linking, the category attribute is a preferred but not required element. This proves, however, that the category is no doubt an essential attribute for the lexicons. In many cases the category will determine the position of the connective, although not always. This is especially true for adverbs, which may be more or less restricted in terms of their position (initial, internal or final in the second argument). But, as mentioned in 4.2.1, sentence positions of the connectives are language-specific and it would be difficult to establish a common descriptive system for all languages in Connective-Lex. Nevertheless, if possible, this would prove very useful for the linking process. Finally, linking might also consider additional information such as mood and register. Although not required, adding this level of granularity would certainly improve the linking. The recommendation for the linking process would then be to consider the levels of word form and sense as the minimum required information, plus the category when applicable.

6. Conclusion

77 Even though connectives do not form a syntactically homogeneous class, they can be meaningfully analyzed by means of the standard repertoire of relations in lexical semantics: we find synonymy (e.g., German *obwohl*, *obzwar*), plesionymy (e.g., *although*, *though*), antonymy (e.g., *if-unless* or *before-after*), hyponymy (e.g., *but*, *although*), and polysemy (e.g., *while*). This is one indication why taking a broad lexical perspective on these items can be productive. Connective-Lex.info is intended as a starting point for collecting information about these lexical items, and for additionally adopting a multilingual perspective and building bridges between languages. While our initial set of common attributes to describe connectives across languages is relatively small, in this paper we have hinted at various possible directions for adding "depth" to the descriptions. For example, we mentioned register preferences (Section 4.2) and mood restrictions (Section 5.4).

78 On balance, we found that the definition of "connective", stemming from the work on German by Pasch et al. (2003), appears to cover what we found in other languages too, and thus functions as our language-neutral working definition. Likewise, the minimal set of attributes for syntax and semantics, derived from a reduced version of the original DiMLex, also appears to be useful for the other languages.

79 From a methodological viewpoint, lexical description and corpus-based research can mutually benefit from each other: as pointed out in Section 2.2, several approaches have been proposed to derive connective lists (which can be the starting point for a

lexicon) from mono- or bilingual corpora (e.g., Versley, 2010); likewise, an existing list or lexicon can be verified and extended by mining distributionally similar words from corpora (e.g., Bourgonje et al., 2017). Moving beyond this initial phase of creating inventories, features of lexical description can be tested with corpora and lexical entries in turn be improved. Finally, we plan to include frequency information into the lexical entries, as one way of reflecting degrees of ubiquity of connectives, their different senses, and possible genre/register dependencies. While we have compiled some information of this kind for the German DiMLex already, it is not straightforward to generalize this across languages: as soon as similar kinds of numbers appear in the different languages of Connective-Lex, one has to be clear whether and in what way they can be meaningfully compared.

80 One central question for extending Connective-Lex is the means of linking the entries across languages. The principal alternatives are establishing pairwise correspondences, or declaring one language as the pivot to which the others are related. Both approaches have their advantages, depending on the use case: precise comparisons of related connectives in two languages (cf. Section 2.2) can yield detailed insights which could considerably improve the usefulness of Connective-Lex for purposes of translation studies, if it is possible to systematically represent such information in a complex linkage scheme. The pivot approach – where presumably English is the most useful candidate – on the other hand would make the system more useful for language learners, since corresponding connectives of any two languages can in principle be found via a commonly understood “interlingua”. Of course, one has to be aware of limitations: the mapping from language *x* to English will be only an approximation; the mapping from English to language *y* will also be approximative; and then the comparison between languages *x* and *y* is bound to be more approximative.

81 For both the direct connections or the pivot connections, as we pointed out in Section 5, the meaningful units for a mapping need to be defined, which is not trivial due to the ambiguity of connectives, and its specific representation in the lexicons. We argued that word form/sense pairs are the most promising units, and we plan to use these in our future work on enhancing the language links in Connective-Lex. The overall goal is to provide more fine-grained correspondences than the current implementation of the search interface does (i.e., the ability to retrieve connectives across languages that share the same sense and/or the same syntactic category).

References

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, L.-M., Le Draoulec, A., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M. & Vieu, L. 2012. An Empirical Resource for Discovering Cognitive Principles of Discourse Organization: The ANNODIS Corpus. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation – LREC 2012*. Luxembourg: European Language Resources Association: 2727-2734. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/836_Paper.pdf.

Al-Saif, A. & Markert, K. 2010. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation – LREC 2010*. Luxembourg: European Language Resources Association: 2046-2053. Available online: http://www.lrec-conf.org/proceedings/lrec2010/pdf/479_Paper.pdf.

Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht – Boston – London: Kluwer Academic Publishers.

Asher, N. & Lascarides, A. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Behrens, B. & Fabricius-Hansen, C. 2002. Connectives in Contrast: A Discourse Semantic Study of Elaboration Based on Corpus Research. In H. Hasselgård, S. Johansson & B. Behrens (eds.), *Information Structure in a Cross-linguistic Perspective*. Amsterdam – New York: Rodopi: 45-61.

Bourgonje, P., Grishina, Y. & Stede, M. 2017. Toward a Bilingual Lexical Database on Connectives: Exploiting a German/Italian Parallel Corpus. In R. Basili, M. Nissim & G. Satta (eds.), *Proceedings of the Fourth Italian Conference on Computational Linguistics – CLIC-IT 2017 (11-12 December 2017, Rome)*. Turin: Accademia University Press: 53-58. Available online: https://books.openedition.org/aaccademia/2360.

Bourgonje, P., Hoek, J., Evers-Vermeul, J., Redeker, G., Sanders, T. & Stede, M. 2018. Constructing a Lexicon of Dutch Discourse Connectives. *Computational Linguistics in the Netherlands Journal* 8: 163-175.

Breindl, E., Volodina, A. & Wassner, U.H. 2014. *Handbuch der deutschen Konnektoren*. Berlin – Munich – Boston: De Gruyter. Vol. 2: *Semantik der deutschen Satzverknüpfer*.

Briz, A., Pons, S. & Portolés, J. (eds.) 2008. *Diccionario de partículas discursivas del español*. URL: http://www.dpde.es.

Colinet, M. 2015. *Report for the STSM (Potsdam 2015)*. Université Paris Diderot – Paris 7. 1-10. Available online: http://textlink.ii.metu.edu.tr/files/stsmreports/STSM2report-Colinet-DimLexConn.pdf.

Danlos, L. 2006. "Discourse Verbs" and Discourse Periphrastic Links. In M. Butt (ed.), *Proceedings of KONVENS 2006 (Konferenz zur Verarbeitung natürlicher Sprache), Universität Konstanz*. University of Konstanz: Konstanz Online Publication System (KOPS): 160-166. Available online: http://www.ub.uni-konstanz.de/kops/volltexte/2006/2013/.

DANLOS, L., RYSOVÁ, K., RYSOVÁ, M. & STEDE, M. 2018. Primary and Secondary Discourse Connectives: Definitions and Lexicons. *Dialogue and Discourse* 9 (1): 50-78. Available online: http://dad.uni-bielefeld.de/index.php/dad/article/download/3734/3617.

DAS, D., SCHEFFLER, T., BOURGONJE, P. & STEDE, M. 2018. Constructing a Lexicon of English Discourse Connectives. In K. KOMATANI, D. LITMAN, K. YU, A. PAPANGELIS, L. CAVEDON & M. NAKANO (eds.), *Proceedings of the 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue – SIGDIAL 2018 (12-14 July 2018, Melbourne, Australia)*. Stroudsburg: Association for Computational Linguistics: 360-365. Available online: https://www.aclweb.org/anthology/W18-5042.

DOMBEK, F. 2017. Connective-Lex.info: A Web App for a Multilingual Connective Database. Bachelor thesis. University of Potsdam, Department of Linguistics. Available online: https://github.com/discourse-lab/Connective-Lex.info/blob/master/Connective-Lex.info%20-%20Bachelor%20thesis.pdf.

FELTRACCO, A., JEZEK, E., MAGNINI, B. & STEDE, M. 2016. LICO: A Lexicon of Italian Connectives. In A. CORAZZA, S. MONTEMAGNI & G. SEMERARO (eds.), *Proceedings of the Third Italian Conference on Computational Linguistics – CLIC-IT 2016 (5-6 December 2016, Napoli)*. Turin: Accademia University Press: 141-145. Available online: https://books.openedition.org/aaccademia/1770.

GROTE, B., LENKE, N. & STEDE, M. 1997. Ma(r)king Concessions in English and German. *Discourse Processes* 24 (1): 87-117.

HAJLAOUI, N. & POPESCU-BELIS, A. 2012. Translating English Discourse Connectives into Arabic: A Corpus-Based Analysis and an Evaluation Metric. In A. FARGHALY & F. OROUMCHIAN (eds.), *Proceedings of the Fourth Workshop on Computational Approaches to Arabic Script-Based Languages at the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012, San Diego, CA, USA, November 1, 2012)*. 1-8. Available online: http://www.mt-archive.info/AMTA-2012-WS-Arabic.pdf.

HANNEFORTH, T., HEINTZE, S. & STEDE, M. 2003. Rhetorical Parsing with Underspecification and Forests. In *Companion Volume of the Proceedings of HLT-NAACL 2003 – Short Papers*. Stroudsburg: Association for Computational Linguistics: 31-33. Available online: https://www.aclweb.org/anthology/N03-2011.

HOEK, J., ZUFFEREY, S., EVERS-VERMEUL, J. & SANDERS, T.J.M. 2017. Cognitive Complexity and the Linguistic Marking of Coherence Relations. A Parallel Corpus Study. *Journal of Pragmatics* 121: 113-131.

KESKES, I., BENAMARA ZITOUNE, F. & BELGUITH, L.H. 2014. Learning Explicit and Implicit Arabic Discourse Relations. *Journal of King Saud University – Computer and Information Sciences* 26 (4): 398-416.

KNOTT, A. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD dissertation. University of Edinburgh.

KNOTT, A. & DALE, R. 1994. Using Linguistic Phenomena to Motivate a Set of Coherence Relations. *Discourse Processes* 18 (1): 35-62.

KNOTT, A. & SANDERS, T.J.M. 1998. The Classification of Coherence Relations and Their Linguistic Markers: An Exploration of Two Languages. *Journal of Pragmatics* 30 (2): 135-175.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit (September 12-16, 2005, Phuket, Thailand) – MT Summit X*. 79-86. Available online: http://www.mt-archive.info/MTS-2005-Koehn.pdf.

Laali, M. & Kosseim, L. 2014. Inducing Discourse Connectives from Parallel Text. In J. Tsujii & J. Hajic (eds.), *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Stroudsburg: Association for Computational Linguistics: 610-619. Available online: https://www.aclweb.org/anthology/C14-1058.

Lin, Z., Ng, H.T. & Kan, M.-Y. 2014. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering* 20 (2): 151-184.

Mann, W.C. & Thompson, S.A. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8 (3): 243-281.

Mazeikiene, V. & Vaiciuniene, V. 2016. Cross-linguistic Variation of the Discourse Marker “However” in Parallel Corpora and Translation of “However” from English into Lithuanian. *Language in Different Contexts* 7 (1): 142-152.

Mendes, A. & Río, I. del 2018. Using a Discourse Bank and a Lexicon for the Automatic Identification of Discourse Connectives. In A. Villavicencio, M. Viviane, A. Abad, H. Caseli, P. Gamallo, C. Ramisch, H.R. Gonçalo Oliveira & G.H. Paetzold (eds.), *Computational Processing of the Portuguese Language. 13th International Conference, PROPOR 2018, Canela, Brazil, September 24-26, 2018, Proceedings*. Cham: Springer International Publishing: 211-221.

Mendes, A., Río, I. del, Stede, M. & Dombek, F. 2018. A Lexicon of Discourse Markers for Portuguese – LDM-PT. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga, S. Goggi & H. Mazo (eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation – LREC 2018*. Luxembourg: European Language Resources Association: 4379-4384. Available online: https://www.aclweb.org/anthology/L18-1693.

Meurers, D. & Dickinson, M. 2017. Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics. *Language Learning* 67 (S1): 66-95.

Meyer, T. & Poláková, L. 2013. Machine Translation with Many Manually Labeled Discourse Connectives. In B. Webber, A. Popescu-Belis, K. Markert & J. Tiedemann (eds.), *Proceedings of the Workshop on Discourse in Machine Translation – DiscoMT (Sofia, Bulgaria, August 9, 2013)*. Stroudsburg: Association for Computational Linguistics: 43-50. Available online: https://www.aclweb.org/anthology/W13-3306.

Meyer, T., Popescu-Belis, A., Zufferey, S. & Cartoni, B. 2011. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In J.Y. Chai, J.D. Moore, R.J. Passonneau & D.R. Traum (eds.), *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue – SIGDIAL 2011 (June 17-18, 2011, Oregon Science and Health University, Portland, Oregon, USA)*. Stroudsburg: Association for Computational Linguistics: 194-203. Available online: https://www.aclweb.org/anthology/W11-2022.

Mírovský, J., Synková, P., Rysová, M. & Poláková, L. 2017. CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics* 109 (1): 61-91. Available online: https://ufal.mff.cuni.cz/pbml/109/art-mirovsky-et-al.pdf.

Mortier, L. & Degand, L. 2009. Adversative Discourse Markers in Contrast – The Need for a Combined Corpus Approach. *International Journal of Corpus Linguistics* 14 (3): 338-366.

Nedoluzhko, A. & Lapshinova-Koltunski, E. 2018. Pronominal Adverbs in German and Their Equivalents in English, Czech and Russian: Evidence from the Parallel Corpus. In V.P. Selegey (ed.), *Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue" (Moscow, May 30-June 2, 2018)*. 522-532. Available online: http://www.dialog-21.ru/media/4560/_-dialog2018scopus.pdf.

Pasch, R., Brausse, U., Breindl, E. & Wassner, U.H. 2003. *Handbuch der deutschen Konnektoren*. Berlin – New York: De Gruyter.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B.L. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation – LREC 2008*. Luxembourg: European Language Resources Association: 2961-2968. Available online: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

Reitter, D. 2003. Simple Signals for Complex Rhetorics: On Rhetorical Analysis with Rich-Feature Support Vector Models. *LDV-Forum* 18 (1-2): 38-52.

Rösner, D. & Stede, M. 1994. Generating Multilingual Documents from a Knowledge Base: The TECHDOC Project. In M. Nagao & Y. Wilks (eds.), *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 339-343. Available online: https://www.aclweb.org/anthology/C94-1055.

Roze, C., Danlos, L. & Muller, P. 2012. LEXCONN: A French Lexicon of Discourse Connectives. *Discours* 10: 1-15. Available online: https://journals.openedition.org/discours/8645.

Sanders, T.J.M., Spooren, W.P.M. & Noordman, L.G.M. 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes* 15 (1): 1-35.

Scheffler, T. 2014. A German Twitter Snapshot. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation – LREC 2014*. Luxembourg: European Language Resources Association: 2284-2289. Available online: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1146_Paper.pdf.

Scheffler, T. & Stede, M. 2016. Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation – LREC 2016*. Luxembourg: European Language Resources Association: 1008-1013. Available online: http://www.lrec-conf.org/proceedings/lrec2016/pdf/274_Paper.pdf.

Scheffler, T., Stede, M., Bourgonje, P. & Dombek, F. 2018. A Multilingual Database of Connectives: Connective-Lex.info. In L.-M. Ho-Dac & P. Muller (eds.), *Cross-linguistic Discourse Annotation: Applications and Perspectives (TextLink2018 – Final Action Conference; March 19-21, 2018, Toulouse, France)*. 144-150. Available online: http://textlink.ii.metu.edu.tr/sites/default/files/textlink_proceedings.pdf.

Stede, M. 2002. DiMLex: A Lexical Approach to Discourse Markers. In A. Lenci & V. Di Tomaso (eds.), *Exploring the Lexicon – Theory and Computation*. Alessandria: Edizioni dell'Orso: 1-15.

Stede, M. 2014. Resolving Connective Ambiguity: A Prerequisite for Discourse Parsing. In H. Gruber & G. Redeker (eds.), *The Pragmatics of Discourse Coherence: Theories and Applications*. Amsterdam – Philadelphia: J. Benjamins: 121-141.

Stede, M. & Heintze, S. 2004. Machine-Assisted Rhetorical Structure Annotation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 425-431. Available online: https://www.aclweb.org/anthology/C04-1061.

Stede, M. & Neumann, A. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation – LREC 2014*. Luxembourg: European Language Resources Association: 925-929. Available online: http://www.lrec-conf.org/proceedings/lrec2014/pdf/579_Paper.pdf.

Stede, M. & Umbach, C. 1998. DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 1238-1242. Available online: https://www.aclweb.org/anthology/C98-2197.

Sweetser, E. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge – New York – Melbourne: Cambridge University Press.

Versley, Y. 2010. Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection. In L. Ahrenberg, J. Tiedemann & M. Volk (eds.), *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*. Tartu: Northern European Association for Language Technology: 83-92.

Webber, B.L., Prasad, R., Lee, A. & Joshi, A. 2016. A Discourse-Annotated Corpus of Conjoined VPs. In A. Friedrich & K. Tomanek (eds.), *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016)*. Stroudsburg: Association for Computational Linguistics: 22-31. Available online: https://www.aclweb.org/anthology/W16-1704.

Webber, B.L., Prasad, R., Lee, A. & Joshi, A. 2019. *The Penn Discourse Treebank 3.0 Annotation Manual*. Philadelphia: University of Pennsylvania. Available online: https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.

Zeyrek, D., Mendes, A. & Kurfali, M. 2018. Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga, S. Goggi & H. Mazo (eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation – LREC 2018*. Luxembourg: European Language Resources Association: 1913-1919. Available online: https://www.aclweb.org/anthology/L18-1301.

Zhou, Y. & Xue, N. 2012. PDTB-Style Discourse Annotation of Chinese Text. In H. Li, C.-Y. Lin, M. Osborne, G.G. Lee & J.C. Park (eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg: Association for Computational Linguistics: 69-77. Available online: https://www.aclweb.org/anthology/P12-1008.

Zhou, Y. & Xue, N. 2015. The Chinese Discourse TreeBank: A Chinese Corpus Annotated with Discourse Relations. *Language Resources and Evaluation* 49 (2): 397-431.

Zufferey, S. 2016. Discourse Connectives across Languages. Factors Influencing Their Explicit or Implicit Translation. *Languages in Contrast* 16 (2): 264-279.

Zufferey, S. & Cartoni, B. 2012. English and French Causal Connectives in Contrast. *Languages in Contrast* 12 (2): 232-250.

Appendix

PDTB2 relation	PDTB3 relation
Comparison	COMPARISON
Comparison.Concession Comparison.Concession.Contra-expectation Comparison.Concession.Expectation	COMPARISON:Concession COMPARISON:Concession:Arg2-as-denier COMPARISON:Concession:Arg1-as-denier
Comparison.Contrast Comparison.Contrast.Juxtaposition Comparison.Contrast.Opposition	COMPARISON:Contrast
Comparison.Pragmatic contrast	COMPARISON:Contrast+SpeechAct
Comparison.Pragmatic concession	COMPARISON:Concession+SpeechAct
Contingency	CONTINGENCY
Contingency.Cause.Reason	CONTINGENCY:Cause:Reason
Contingency.Cause.Result	CONTINGENCY:Cause:Result
Contingency.Pragmatic cause.Justification	CONTINGENCY:Cause+Belief:Reason+Belief
Contingency.Condition Contingency.Condition.Factual past Contingency.Condition.Factual present Contingency.Condition.General Contingency.Condition.Hypothetical Contingency.Condition.Unreal past Contingency.Condition.Unreal present	CONTINGENCY:Condition
Contingency.Pragmatic condition. Implicit assert. Contingency.Pragmatic condition. Relevance	CONTINGENCY:Condition+SpeechAct
Expansion	EXPANSION
Expansion.Alternative Expansion.Alternative.Conjunctive Expansion.Alternative.Disjunctive	EXPANSION:Disjunction
Expansion.Alternative.Chosen alternative	EXPANSION:Substitution

Expansion.Conjunction	EXPANSION:Conjunction
Expansion.Exception	EXPANSION:Exception
Expansion.Instantiation	EXPANSION:Instantiation
Expansion.List	EXPANSION:Conjunction
Expansion.Restatement	EXPANSION:Level-of-detail
Expansion.Restatement.Equivalence	EXPANSION:Equivalence
Expansion.Restatement.Generalization	EXPANSION:Level-of-detail:Arg1-as-detail
Expansion.Restatement.Specification	EXPANSION:Level-of-detail:Arg2-as-detail
Temporal	TEMPORAL
Temporal.Asynchronous	TEMPORAL:Asynchronous
Temporal.Asynchronous.Precedence	TEMPORAL:Asynchronous:Precedence
Temporal.Asynchronous.Succession	TEMPORAL:Asynchronous:Succession

Table 1 – Mapping from PDTB2 to PDTB3 connective senses used in Connective-Lex. Note that not all PDTB2 or PDTB3 senses can be automatically linked (some completely new relations were added in PDTB3). In addition, in some cases, only level 2 relations (e.g., “Condition” or “Exception”) can be linked to automatically, because the directionality of the relation is not systematically encoded in the PDTB2 relations

SDRT relation	PDTB3 relation
alternation	EXPANSION:Disjunction
background	TEMPORAL:Asynchronous:Precedence
background-inverse	TEMPORAL:Asynchronous:Succession
concession	COMPARISON:Concession
condition	CONTINGENCY:Condition
consequence	CONTINGENCY:Condition
continuation	EXPANSION:Conjunction
contrast	COMPARISON:Contrast
detachment	EXPANSION:Exception
digression	EXPANSION:Conjunction
elaboration	EXPANSION:Level-of-detail
evidence	EXPANSION:Conjunction
explanation	EXPANSION:Manner EXPANSION:Level-of-detail CONTINGENCY:Cause:Reason
explanation*	CONTINGENCY:Cause+Belief
flashback	TEMPORAL:Asynchronous:Succession
goal	CONTINGENCY:Purpose
narration	TEMPORAL:Asynchronous:Precedence EXPANSION:Conjunction
parallel	COMPARISON:Similarity
rephrasing	EXPANSION:Equivalence
result	CONTINGENCY:Cause:Result
result*	CONTINGENCY:Cause:Result+Belief CONTINGENCY:Cause:Result+SpeechAct
summary	EXPANSION:Level-of-detail:Arg2-as-detail
temploc	TEMPORAL
violation	EXPANSION:Exception

Table 2 – Mapping from SDRT to PDTB3 connective senses used in Connective-Lex