Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1128–1134
Marseille, 11–16 May 2020

CzeDLex 0.6 and its Representation in the PML-TQ

Jiří Mírovský, Lucie Poláková and Pavlína Synková
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{mirovsky, polakova, synkova}@ufal.mff.cuni.cz

Abstract

CzeDLex is an electronic lexicon of Czech discourse connectives with its data coming from a large treebank annotated with discourse relations. Its new version CzeDLex 0.6 (as compared with the previous version 0.5, which was published in 2017) is significantly larger with respect to manually processed entries. Also, its structure has been modified to allow for primary connectives to appear with multiple entries for a single discourse sense. The lexicon comes in several formats, being both human and machine readable, and is available for searching in PML Tree Query, a user-friendly and powerful search tool for all kinds of linguistically annotated treebanks. The main purpose of this paper/demo is to present the new version of the lexicon and to demonstrate possibilities of mining various types of information from the lexicon using PML Tree Query; we present several examples of search queries over the lexicon data along with their results. The new version of the lexicon, CzeDLex 0.6, is available on-line and was officially released in December 2019 under the Creative Commons License.

Keywords: lexicon, discourse connectives, Czech, searching, PML-TQ

1. Introduction and Outline

Despite a strong emphasis on the use of neural networks also in such areas as discourse parsing (Wang et al., 2015; Hooda and Kosseim, 2017; Knaebel et al., 2019), development of electronic lexicons of discourse connectives has been experiencing an unprecedented boom in the last years as well. Inspired by first large-scale electronic lexicons of discourse connectives, most importantly by an XML-based and machine readable DiMLex for German (Stede, 2002; Scheffler and Stede, 2016), and also by the more human-oriented DPDE, a dictionary of Spanish discourse markers (Briz et al., 2003), quite many more lexicons of discourse connectives for other languages have emerged recently: LexConn for French (Roze et al., 2012), LICO for Italian (Feltracco et al., 2016), CzeDLex for Czech (Mírovský et al., 2017a), DiMLex-Eng for English (Das et al., 2018), LDM-PT for Portuguese (Mendes et al., 2018), and others. Most of these resources have been gradually unified in Connective-Lex (Stede et al., 2019), a multi-language database of discourse connectives currently covering 9 languages.[1]

CzeDLex is an annotated electronic lexicon that provides semantic and morpho-syntactic information about Czech discourse connectives. Its first version, CzeDLex 0.5 (Mírovský et al., 2017a), was released in 2017 and described in Mírovský et al. (2017b). In the present contribution, after a brief overview of the lexicon properties in Section 2, we introduce CzeDLex 0.6, an updated version of the lexicon that contains significantly larger amount of manually processed entries and also redefines the basic structure of the lexicon, to comply with results of the recent research on the topic (Section 3). The lexicon can be accessed in several ways: browsed on-line as HTML web pages, locally in its native XML format via the tree editor TrEd, etc. In Section 4, which represents the core of the paper for the demo, we focus on the possibility to search in the lexicon using PML-Tree Query, a powerful search engine for treebanks, and present several examples of linguistically motivated queries. The final section (Section 5) gives information on the availability of the lexicon and future plans.

2. CzeDLex

Lexicons of discourse connectives can be built in several ways: manually by consulting existing printed lexicons or other resources, by translating from a lexicon from another language, and finally, by exploiting existing discourse-annotated corpora. CzeDLex is based on manual annotation of more than 21 thousand explicit discourse relations in a large corpus of Czech newspaper texts, Prague Discourse Treebank 2.0 (PDiT 2.0, Rysová et al. (2016)), which we call 'the source corpus' in the subsequent text. Annotation of discourse relations in the source corpus covers explicit discourse relations, i.e. discourse relations anchored by a surface-present connective, with a modified version of the PDTB 2.0 taxonomy for the annotation of senses (Poláková et al., 2013).

Using the data-driven approach in connection with the large source corpus means that the lexicon contains not only the 'typical' readings of connectives but also less frequent and rare ones, which are often linguistically more interesting than the common cases; each of them is in the lexicon supported by real text examples. Corpus frequencies, which are also part of the lexicon, reflect and help distinguish the central, less typical and marginal ways the connectives are used in language, as we demonstrate in several example queries in Section 4.[2]

The process of extracting connectives and their properties from the source corpus was described in Synková et al. (2017). An automatic script extracted information about

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

[2] The lexicon also reflects and quantifies non-connective readings of polyfunctional connective-like expressions.

the annotated discourse relations and grouped occurrences of connective variants (e.g., *teda* is an informal variant of *tedy* [*so*]), complex forms of connectives[3] (e.g., *a také* [*and also*]) and modifications[4] (e.g., *právě protože* [*exactly because*]). Subsequently, manual checks and additions were performed on the most frequent entries, resulting into the publication of CzeDLex 0.5, with (only) 18 out of 205 entries fully manually checked and enriched with additional linguistic information; these 18 entries, however, covered more than 2/3 of all occurrences of discourse relations in the source corpus.

The discourse annotation in the source corpus, and subsequently also CzeDLex, contain both primary and secondary connectives. Primary connectives are short and grammaticalized expressions belonging to certain parts of speech (conjunctions, particles and some types of adverbs), such as *a* [*and*], *zatímco* [*while*], *protože* [*because*], *ovšem* [*however*], etc. Lexicon entries for primary connectives are represented by lemmas of the connectives (which are in most cases the connectives themselves). Secondary connectives are usually multiword phrases such as *z tohoto důvodu* [*for this reason*], *v případě že* [*in case that*], *vzhledem k tomu* [*with respect to that*], *jak dodává* [*as he adds*], i.e. expressions that are not fully grammaticalized and show a high degree of variability (see Rysová and Rysová (2015)). Lexicon entries for secondary connectives are represented by their core words, i.e., for example, *důvod* [*reason*], *případ* [*case*], *vzhledem k* [*with respect to*], *dodat* [*to add*].

The logical structure of CzeDLex was described and the selected choice discussed in detail in Mírovský et al. (2016b). Most importantly, each first-level entry (representing a connective) is further divided into second-level entries that represent individual usages of the connective, i.e. discourse senses expressible by the connective. These usages are complemented with lists of complex forms and modifications, additional information about argument semantics, ordering of the arguments, morpho-syntactic information about the connective and its position with respect to the arguments, examples, corpus counts, etc.

Structuring the lexicon this way, i.e. putting emphasis on discourse senses, reflects the primary use of connectives in language, i.e. to express semantico-pragmatic relations in discourse. In practice, it allows for linking the lexicon to similar lexicons in other languages on the level of discourse senses, significantly narrowing the set of translation candidates for a connective used in a particular sense. Poláková et al. (2020) demonstrate such a procedure for CzeDLex and DiMLex.

3. CzeDLex 0.6

After the automatic extraction of raw lexicon data from the source corpus, manual work was needed to check, add and refine information that (i) the source corpus annotations did not contain,[5] (ii) would be too difficult to obtain automatically,[6] and (iii) the source corpus was not big enough to cover all possibilities.[7] The steps of manual enhancements of the lexicon entries was described in Mírovský et al. (2017b).

Manual work also revealed inconsistencies of the annotation in the source corpus, namely different discourse senses annotated in very similar contexts, contexts with an obvious discourse relation erroneously not annotated (and thus treated as a non-connective usage in the lexicon) or inconsistent treatment of complex forms and modifications in similar contexts. All these phenomena were fixed in the lexicon and listed for correction in a future release of the source corpus.

3.1. Updates in the Structure

During the manual checks, the overall number of lexicon entries changes. The initial automatically extracted lexicon contained connectives in the form of strings, as they were annotated in the PDiT 2.0 data. During the manual checking process, certain strings required to be (i) merged under one lexicon entry, (ii) shifted from one entry to another or (iii) split in order to form individual entries.

(i) Merging lexicon entries included for instance:

- merging two style variants of one connective (with no other difference in their characteristics), e.g. *jenom* as an informal and spoken variant was merged with its canonical form *jen* [*only*];
- merging separately[8] used adverbs (as *poté* [*then*], cf. Example 1) with structures where they, together with specific subordinators, introduce a dependent temporal clause in Czech (as *poté, co* [*after*, lit. *after that, when*], cf. Example 2).

(1) *Jirka dodělal doktorát. Poté se mu nabídky práce jen hrnuly.*

[*George completed his doctoral degree. After that, job offers started pouring in.*]

Example 1 documents a case where the connective word appears in the argument happening later (succession). Only this type of argument ordering is possible.

Example 2 documents both ordering settings for the connective phrase *poté, co* introducing a temporal dependent clause. The connective phrase is present in both arguments, the part *poté* is, again, in the argument happening later (succession).

(2) *Poté, co Jirka dodělal doktorát, se mu nabídky práce jen hrnuly.*

[*After George completed his doctoral degree, job offers started pouring in.*]

[3] A complex form consists of two or more connective words.

[4] modified, e.g. intensified connectives

[5] English translations, Czech synonyms, register and basic syntactic characteristics of secondary connectives

[6] argument semantics, ordering of the arguments and dependency schemes representing various syntactic patterns of secondary connectives

[7] integration (a position of the connective in respect with the arguments)

[8] mostly inter-sententially

Cítila jeho vůni gin & tonik ještě poté, co odešel.

[*She could smell his gin & tonic scent even after he left.*]

Interestingly, a different placement of comma in Example 3[9] puts the typical connective of succession into the argument expressing precedence (the whole *poté co*-phrase). Such cases must be clearly and carefully distinguished in order not to mislead any discourse parsing system – differences in argument ordering possibilities imply differences in where to look for the arguments of the connective, the placement of the same connective word/phrase in one or the other argument of the same relation is a similar case.

(3) *Prospal celé odpoledne, poté co odmítl cokoliv solidního pozřít.*

[*He slept all afternoon after he refused to eat anything solid.*]

These patterns motivated an enhancement in the lexicon structure: it newly allows existence of several identical semantic relations in one entry of a primary connective, e.g. 'precedence-succession-1' and 'precedence-succession-2' etc., with differences in additional attributes (ordering, synt. structure, argument semantics – direction of an asymmetric relation). Such a change in the lexicon structure resembles the entry structure for secondary connectives – typically multiword phrases with various syntactic realizations for one discourse meaning.

(ii) Moving items from one entry to another can be exemplified by the archaic expression *neb*, which can represent an abbreviated form of either *neboť* [because] or *nebo* [or]. The automatic lemmatization of the source corpus did not disambiguate these rare occurrences correctly and the error projected all the way up to discourse annotation.

(iii) Splitting entries/introduction of new first-level entries was applied e.g. for:

- the connective *přece jen* [*after all*, lit. *yet only*], which was originally (as a complex form) a part of both the entries *přece* and *jen*; the manual check revealed that when co-occurring, these forms express a concessive meaning which is not possible for the individual forms occurring separately;

- forms *anebo* and *aneb* [in English both translated as *or*], which were originally wrongly merged for its formal similarity and the same historical origin. Nevertheless, these forms are not interchangeable in modern Czech: the connective *anebo* [*or*] expresses an alternative, whereas the connective *aneb* [*in other words*] expresses 'equivalence' only.

4. Searching in the Data

PML-Tree Query (PML-TQ) is a powerful client–server system for searching in linguistically annotated treebanks (Pajas and Štěpánek, 2009). It is a general tool that can be used for any treebank encoded in Prague Markup Language (see examples of using the PML-TQ for various treebanks in Štěpánek and Pajas (2010), Mírovský et al. (2016a), Onambélé et al. (2017)).

Prague Markup Language (PML; Hana and Štěpánek (2012))[10] is an abstract XML format designed for complex linguistic annotation of treebanks, technically of any text data represented as tree structures (i.e., also lexicons). Data in the PML format can be browsed and edited in TrEd, a highly customizable tree editor (Pajas and Štěpánek, 2008).[11] The PML and TrEd, together with the PML-Tree Query system, form a general framework for treebank annotation and data processing.

CzeDLex has been developed in this framework and as such, it can be directly searched using the PML-TQ. This section gives examples that demonstrate types of queries that can be run on the lexicon, combining the powerful search engine with the rich annotation of entries in CzeDLex.

Creating a basic query in the PML-TQ actually means to define a tree[12] that should be found in a result tree as a subtree. The query is then processed and either individual results are displayed, or – using so called output filters – all matches of the query are summarized and reported in the form of a table.

Example 1. The tree in Figure 1 represents a query that searches in the lexicon for connectives that are able to express both discourse relations 'synchrony' and 'precedence–succession'. The query tree reflects the lexicon structure: the node `$1` of type `c-lemma` represents the first-level entry in the lexicon, i.e. the whole entry for a connective. Its child, the node of type `c-usages`, represents all connective usages of the connective. It has two children nodes of type `c-usage` representing the two required usages. The output filter (specified after >>) creates a list of such connectives (lexically stored in attribute `text` of the node `$1`), sorted alphabetically.

The textual representation of the query can either be typeset/edited directly or is automatically created by the system from the graphically created version of the query:

```
c-lemma $1 :=
  [ c-usages
    [ c-usage
        [ sense = "synchrony" ],
      c-usage
        [ sense = "precedence-succession" ] ] ];
>> for $1.text give $1 sort by $1
```

Table 1 represents the result of the query, i.e. 11 connectives that match the query, printed and sorted alphabetically

[9] caused possibly by the tendency of such phrases in Czech to gradually form single-word expressions

[10] see http://ufal.mff.cuni.cz/jazz/PML/

[11] TrEd is written in Perl and can be easily customized for a desired purpose by extensions that are included into the system as modules. It is available from https://ufal.mff.cuni.cz/tred/ under the GPL – The General Public Licence.

[12] i.e., draw a tree in the graphical client environment mostly by clicking on buttons and selecting from lists of options

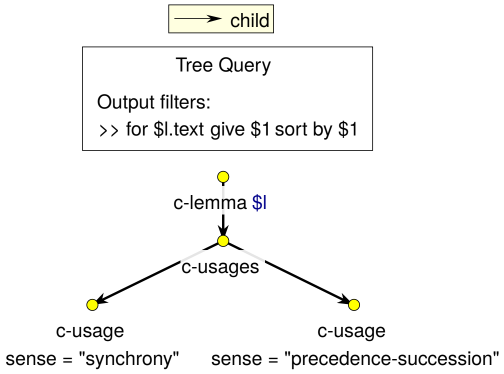


Figure 1: Graphical representation of the query. The output filter is displayed above the query tree. The semantics of the colours of the arrows is explained at the very top of the query (here: black arrows mean the child relation in the tree structure).

by the output filter; approximate English translations have been added here.

a [*and*]
až [*when, as soon as*]
dokud [*until*]
jakmile [*if, as soon as*]
kdy [*when*]
kdykoli [*whenever*]
když [*when*]
mezitím [*in the meantime*]
tím [*by that*]
že [*that*]

Table 1: Result of the query (1): connectives able to represent both 'synchrony' and 'precedence–succession'.

Example 2. The query in the second example with its graphical representation in Figure 2 searches for manually checked primary connectives that can express 'condition'. The output filter summarizes the results and prints the connectives along with number of senses they can express, i.e. (in the descending order) according to their semantic ambiguity.
Table 2 represents the result of the query. Again, approximate English translations have been added here.

Example 3. The query in Figure 3 searches for connectives with symmetric senses[13] and – using an output filter – lists the connectives and their symmetric relations in the descending order according to their counts of occurrencies in the source corpus.
Table 3 represents a sample of the result of the query.

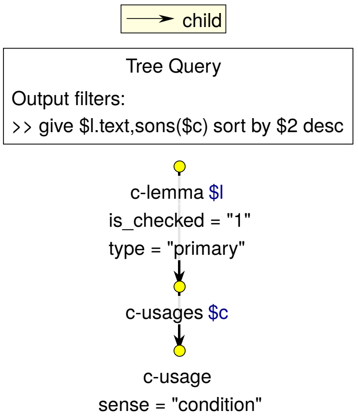


Figure 2: Graphical representation of the query (2).

a [*and*]	11
když [*when*]	10
ale [*but*]	9
tak [*so*]	9
ovšem [*however*]	8
jestliže [*if*]	6
pak [*then*]	6
li [*if*]	5
potom [*then*]	5
pokud [*if*]	4
ať [*no matter how, be it or not*]	2
také [*also*]	2

Table 2: Result of the query (2): manually checked primary connectives able to express 'condition', sorted according to their sense ambiguity.

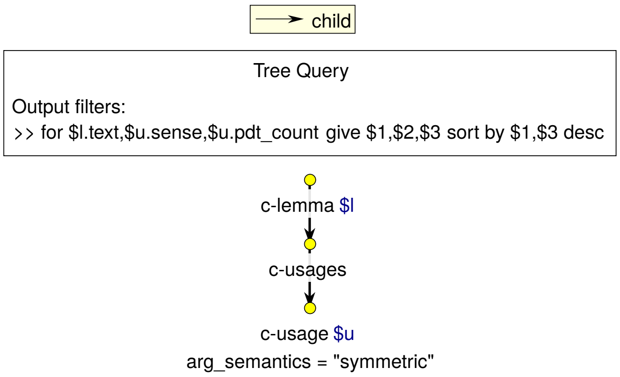

child
Tree Query
Output filters:
>> for $l.text,$u.sense,$u.pdt_count give $1,$2,$3 sort by $1,$3 desc
c-lemma $l
c-usages
c-usage $u
arg_semantics = "symmetric"

Figure 3: Graphical representation of the query (3).

Example 4. The query in Figure 4 lists all connectives in the lexicon according to the ratio of their occurrences in intra-sentential relations in the source corpus.[14] The output filter in this example demonstrates that output filters can be

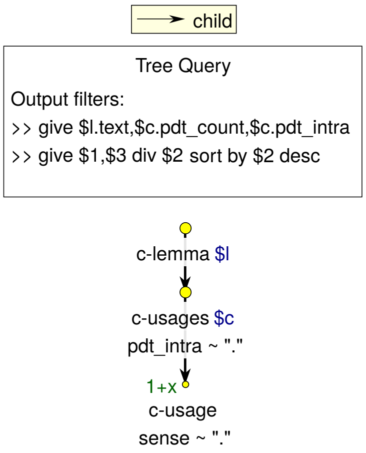
[13] Senses such as 'conjunction' and 'opposition' where the semantics of the arguments is the same for both arguments, as opposed to asymmetric types of relations such as 'reason–result' where the semantics of the arguments differs (one representing the reason, the other the result).

[14] Technically, we take only those lexicon entries that have information about the number of all intra-sentential relations at the node representing all connective usages of the connective (c-usages); the bottom node ensures that there is at least one sense present, i.e. the parent c-usages node is not the parent node for non-connective usages.

a [*and*]	conjunction	5948
a	confrontation	32
a	opposition	30
a	synchrony	18
a	equivalence	9
aby [*(in order) to*]	conjunction	5
ale [*but*]	opposition	1258
ale	confrontation	49
ale	conjunction	22
ale	pragmatic contrast	17
alespoň [*at least*]	disjunctive alternative	3
alespoň	opposition	1
aneb [*or*]	equivalence	3
anebo [*or*]	disjunctive alternative	27
anebo	conjunctive alternative	4
ani [*nor, not even*]	conjunction	57
ani	opposition	2

Table 3: A sample of the result of the query (3): connectives and their symmetric relations sorted according to counts of their occurrences in the source corpus. Approximate English translations have been added.

put one after another – the subsequent filter is applied on the output of the preceding filter.



Output filters:

```
>> give $l.text,$c.pdt_count,$c.pdt_intra
>> give $1,$3 div $2 sort by $2 desc
```

Figure 4: Graphical representation of the query (4).

ač [*although*]	1
ať [*no matter how, be it or not*]	1
...	
pokud [*if*]	0.99
neboť [*because*]	0.99
aby [*(in order) to*]	0.99
...	
čili [*or, in other words*]	0.5
kvůli [*because of*]	0.5
na rozdíl [*in contrast*]	0.5
nehledě na [*regardless of*]	0.5
...	
dále [*also, subsequently*]	0.07
upřesnit [*to specify*]	0.06
souvislost [*respect, connection*]	0.05
ani + případ [*not even in such case*]	0
i potom [*even then*]	0
přece jen [*nonetheless*]	0

Table 4: A sample of the result of the query (4): connectives and ratios of their intra-sentential occurrences in the source corpus. (For typographical reasons, decimal numbers have been shortened.)

Table 4 represents a sample of the result of the query. Connectives in the top lines appeared in the corpus only in intra-sentential relations, i.e. the ratio of the intra-sentential relations among all relations is 1. On the other hand, ratio 0 at the bottom lines of the table means that those connectives appeared exclusively inter-sententially in the source corpus.

Example 5. The last example query (Figure 5) combines several pieces of information available in the lexicon to find intra-sentential examples of rare usages of connectives. The output filter proceeds in the following steps (line by line):

1. For each connective and its usage, it extracts essential information from all three matching query nodes: (i) the connective, (ii) the discourse sense, (iii) the corpus count of the usage of the connective divided by the total count of connective usages of the connective, and two pieces of information about corpus text examples for this connective and sense: (iv) information whether the example is intra-sentential or inter-sentential, and (v) the example itself.
2. It selects only results where the ratio of counts of the usage of the connective among all its connective usages is smaller than 1%.
3. It selects only results that are documented by an intra-sentential example.
4. It gets rid of columns that are no more needed and counts (ranks) examples for each pair of a connective and a sense.
5. For each pair of a connective and a sense, it selects only two first examples (one if there are no more).
6. It gets rid of the remaining supporting columns and selects the final pieces of information – a connective, a sense and an example.

Table 5 represents a sample of the result of the query, namely 6 out of 86 total corpus examples of such rare phenomena captured in the lexicon.

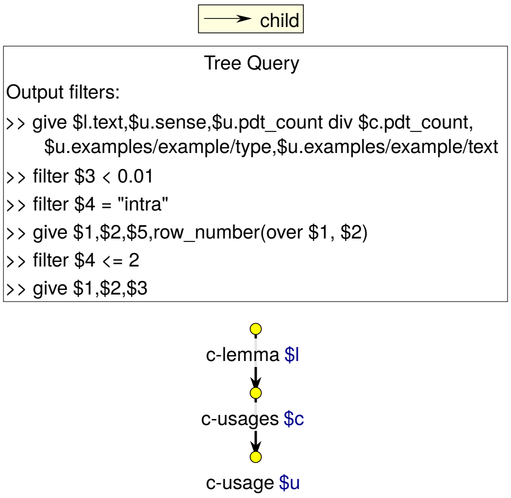
5. Conclusion

We have described updates and changes in the development of CzeDLex, an annotated electronic lexicon of Czech discourse connectives, both in the structure of the lexicon and the size of manually processed entries. The development version of CzeDLex is available on-line as HTML web pages.[15] CzeDLex 0.6 (in both XML and HTML ver-

[15] `http://ufal.mff.cuni.cz/czedlex/`

aby [*(in order) to*]	precedence-succession	Českou republiku opustí zítra, *aby* pokračovala do Rakouska, Moldávie a Zakavkazska a do Moskvy. [She is leaving the Czech Republic tomorrow, *to* continue to Austria, Moldova and the Transcaucasus, and to Moscow.]
aby	precedence-succession	Patří sem zejména cyklické vlny nejrůznějších afér, které vzplanou jasným žárem, *aby* o několik dní později stejně rychle zhasly. [These include, in particular, cyclical waves of all sorts of affairs that flare up with bright passion *to* fade out quickly a few days later anyway.]
když [*when*]	explication	Šťastnější byli nakonec slávisté, *když* v poslední minutě dal branku Jakovenko. [Supporters of Slávia were luckier in the end, *when* Jakovenko scored in the last minute.]
když	explication	A tak v podstatě jedinými, kdo sebe ani ostatní neblamují, jsou bosenští Srbové a Chorvaté, *když* se mezi sebou dohodli na zřízení tří etnických ministátů. [And so, basically the only ones who do not make fools of themselves or the others, are the Bosnian Serbs and the Croats *when* they agreed among themselves to establish three ethnic ministates.]
přestože [although]	pragmatic contrast	*Přestože* si již nyní mnoho nájemníků upravovalo své byty z vlastních prostředků, zákon až dosud umožňoval, aby jim opravy dražší padesáti korun zaplatil majitel bytu. [*Although* many tenants have already in the past adapted their flats using their own resources, until now the law has allowed them to ask the owner of the flat to pay for repairs more expensive than fifty crowns.]
přestože [although]	confrontation	*Přestože* některé funkce totalitního státu již skončily, jiné - např. živnostenská agenda - se objevily. [*Although* some functions of the totalitarian state have ended, others - such as the trade agenda - have emerged.]

Table 5: A sample of the result of the query (5): up to two intra-sentential examples of rare usages of connectives. English translations have been added here.



```
Output filters:
>> give $l.text,$u.sense,$u.pdt_count div $c.pdt_count,
     $u.examples/example/type,$u.examples/example/text
>> filter $3 < 0.01
>> filter $4 = "intra"
>> give $1,$2,$5,row_number(over $1, $2)
>> filter $4 <= 2
>> give $1,$2,$3
```

Figure 5: Graphical representation of the query (5).

sions) was officially released in December 2019 at the Lindat/Clarin repository,[16] freely available under the Creative Commons License, and is also available on-line as HTML web pages.[17] It contains 205 connectives,[18] with 80 of them fully manually checked and enriched with additional linguistic information. These 80 connectives cover over 90% of all discourse relations annotated in the source corpus, the Prague Discourse Treebank 2.0. In the section dedicated to searching in the lexicon, we have demonstrated possibilities that a combination of a powerful search engine with a richly annotated lexicon of discourse connectives offers to language researchers.

Our plans for further development include the aim to define more precisely the distinction of primary and secondary connectives and better reflect their similarities and differences in the lexicon structure. The categories of complex forms and modifications also exhibit some unclear cases and better distinctive criteria are needed. In accordance with the project funding, we plan to complete manual checks and additions in the whole lexicon by the end of 2021 and release the resulting data as CzeDLex 1.0 under the same license as the previous versions.

6. Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project GA19-03490S). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2018101).

7. References

Briz, A., Bordería, S. P., and Portolés, J. (2003). *Diccionario de partículas discursivas del español*. Data/software, www.dpde.es. Online since 2003.

Das, D., Scheffler, T., Bourgonje, P., and Stede, M. (2018). Constructing a Lexicon of English Discourse Connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365.

Feltracco, A., Jezek, E., Magnini, B., and Stede, M. (2016). LICO: A Lexicon of Italian Connectives. *CLiC it*, page 141.

[16] http://hdl.handle.net/11234/1-3074

[17] http://ufal.mff.cuni.cz/czedlex0.6/

[18] By chance, the total number of connectives in CzeDLex 0.6 is the same as in version 0.5, although many entries have been merged or split.

Hana, J. and Štěpánek, J. (2012). Prague Markup Language Framework. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 12–21, Stroudsburg. Association for Computational Linguistics, Association for Computational Linguistics.

Hooda, S. and Kosseim, L. (2017). Argument Labeling of Explicit Discourse Relations Using LSTM Neural Networks. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 309–315.

Knaebel, R., Stede, M., and Stober, S. (2019). Window-Based Neural Tagging for Shallow Discourse Argument Labeling. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 768–777.

Mendes, A., del Rio, I., Stede, M., and Dombek, F. (2018). A Lexicon of Discourse Markers for Portuguese–LDM-PT. In *11th International Conference on Language Resources and Evaluation*, pages 4379–4384.

Mírovský, J., Poláková, L., and Štěpánek, J. (2016a). Searching in the Penn Discourse Treebank Using the PML-Tree Query. In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1762–1769, Paris, France. European Language Resources Association.

Mírovský, J., Synková, P., Rysová, M., and Poláková, L. (2016b). Designing CzeDLex – A Lexicon of Czech Discourse Connectives. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation*, pages 449–457, Seoul, Korea. Kyung Hee University, Kyung Hee University.

Mírovský, J., Synková, P., Rysová, M., and Poláková, L. (2017a). *CzeDLex 0.5*. Data/Software, Charles University, Prague, Czech Republic.

Mírovský, J., Synková, P., Rysová, M., and Poláková, L. (2017b). CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, (109):61–91.

Onambélé, C., Kopp, M., Passarotti, M., and Mírovský, J. (2017). Converting Latin Treebank Data into an SQL Database for Query Purposes. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, pages 117–122, New York, NY, USA. University of Göttingen, Institute of Computer Science, ACM.

Pajas, P. and Štěpánek, J. (2008). Recent Advances in a Feature-Rich Framework for Treebank Annotation. In Donia Scott et al., editors, *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 2, pages 673–680, Manchester. The Coling 2008 Organizing Committee.

Pajas, P. and Štěpánek, J. (2009). System for Querying Syntactically Annotated Corpora. In Gary Lee et al., editors, *Proceedings of the ACL–IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec. Association for Computational Linguistics.

Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., and Hajičová, E. (2013). Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya. Asian Federation of Natural Language Processing.

Poláková, L., Rysová, K., Rysová, M., and Mírovský, J. (2020). GeCzLex: Lexicon of Czech and German Anaphoric Connectives. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, May. European Language Resources Association (ELRA).

Roze, C., Danlos, L., and Muller, P. (2012). LEXCONN: A French Lexicon of Discourse Connectives. *Discours. Revue de linguistique, psycholinguistique et informatique*, (10).

Rysová, M. and Rysová, K. (2015). Secondary Connectives in the Prague Dependency Treebank. In Eva Hajičová et al., editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 291–299, Uppsala, Sweden. Uppsala University, Uppsala University.

Rysová, M., Synková, P., Mírovský, J., Hajičová, E., Nedoluzhko, A., Ocelák, R., Pergler, J., Poláková, L., Pavlíková, V., Zdeňková, J., and Zikánová, Š. (2016). *Prague Discourse Treebank 2.0*. Data/software, ÚFAL MFF UK, Prague, Czech Republic.

Scheffler, T. and Stede, M. (2016). Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 1008–1013, Portorož, Slovenia. European Language Resources Association (ELRA).

Stede, M., Scheffler, T., and Mendes, A. (2019). Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

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

Štěpánek, J. and Pajas, P. (2010). Querying Diverse Treebanks in a Uniform Way. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1828–1835, Valletta, Malta. European Language Resources Association.

Synková, P., Rysová, M., Poláková, L., and Mírovský, J. (2017). Extracting a Lexicon of Discourse Connectives in Czech from an Annotated Corpus. In Rachel Roxas, editor, *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 232–240, Cebu, Philippines. Computing Society of the Philippines, National University of Philippines.

Wang, L., Hokamp, C., Okita, T., Zhang, X., and Liu, Q. (2015). The DCU Discourse Parser for Connective, Argument Identification and Explicit Sense Classification. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 89–94.