# Extracting a Lexicon of Discourse Connectives in Czech from an Annotated Corpus

**Pavlína Synková, Magdaléna Rysová, Lucie Poláková, Jiří Mírovský**
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Prague, Czech Republic
{synkova|magdalena.rysova|polakova|mirovsky}@ufal.mff.cuni.cz

## Abstract

We discuss a process of exploiting a large corpus manually annotated with discourse relations – the Prague Discourse Treebank 2.0 – to create a lexicon of Czech discourse connectives (CzeDLex). The data format and the data structure of the lexicon are based on a study of similar existing resources and are adapted for a uniform representation of both primary (such as in English *because, therefore*) and secondary connectives (e.g. *for this reason, this is the reason why*). The main principle adopted for nesting entries in the lexicon is a discourse-semantic type expressed by the given connective word, which enables us to deal with a broad formal variability of connectives. We present a technical solution based on the (XML-based) Prague Markup Language that allows for an efficient incorporation of the lexicon into the family of Prague treebanks – it can be directly opened and edited in the tree editor TrEd, processed from the command line in btred, interlinked with its source corpus and queried in the PML-Tree Query engine – and also for interconnecting CzeDLex with existing lexicons in other languages.

## 1 Introduction

Recent years witnessed a vivid development of corpora annotated with discourse relations. In connection with this development, electronic lexicons of discourse connectives began to be built, although they are so far much less common. These lexicons present an important source not only for theoretical research of text coherence but they may be also helpful in NLP tasks such as discourse parsing (disambiguation of connective and non-connective usages, determining the semantic type of discourse relations), machine translation, text generation and information extraction. This paper presents the process of developing an electronic lexicon of Czech discourse connectives. The chosen approach is inspired by existing electronic lexicons – most of all by DiMLex (Stede, 2002; Scheffler and Stede, 2016), and also by LexConn (Roze et al., 2012), XML-based inventories of discourse connectives for German and French, respectively, and it follows the theoretical framework for designing a lexicon of discourse connectives outlined in Mírovský et al. (2016b).

The text of this paper is organized as follows: Section 2 presents the discourse-annotated treebank used as the source data for the lexicon, in Section 3, the structure of the lexicon and properties of its entries are described, and CzeDLex is also compared to (mostly) DiMLex. Section 4 describes technical aspects of the lexicon development, including the data format and the automatic extraction of connective properties from the treebank data, and also mentions necessary automatic and manual post-processing steps.

## 2 Prague Discourse Treebank 2.0

The Prague Discourse Treebank 2.0 is built upon the data of the Prague Dependency Treebank (Hajič et al., 2006; Bejček et al., 2013), which is a richly annotated corpus with a multilayer annotation of approx. 50 thousand sentences of Czech journalistic texts. The Prague Dependency Treebank con-

*31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31)*, pages 232–240
Cebu City, Philippines, November 16-18, 2017

| | PDiT 1.0<br>(2012) | PDT 3.0<br>(2013) | PDiT 2.0<br>(2016) |
|---|---|---|---|
| Primary connectives | yes | updated | updated |
| Second relations | | yes | updated |
| Secondary connectives | | | yes |

Table 1: Major changes in the annotation of discourse relations in various published versions of the data.

tains morphological information on each token and two layers of syntactic annotation for each sentence (shallow and deep structure), both layers are represented by dependency trees. Besides, there is an annotation of information structure, coreference, bridging anaphora and multiword expressions. Annotation of discourse relations was carried out on top of deep-syntactic trees (on the so called tectogrammatical layer, see Example 1 and Figure 1) and covers relations expressed by a surface-present connective. A connective is defined as a predicate of a binary relation opening two positions for two text spans as its arguments and signalling a semantic or pragmatic relation between them (compare Prasad et al., 2008). The set of discourse types is inspired by the Penn Discourse Treebank 2.0 sense hierarchy (Prasad et al., 2008) and syntactico-semantic labels used for representation of compound sentences on the tectogrammatical layer (the complete set can be found in Zikánová et al., 2015). The annotation reflects a division of connectives into primary and secondary ones (the terms established by Rysová and Rysová, 2014) and it had two phases – in the first one, primary connectives (i.e. grammaticalized expressions such as *because* or *therefore*) were captured, taking into account only those that anchored relations between arguments containing finite verb forms (Poláková et al., 2013). The second phase covered secondary connectives (i.e. not yet fully grammaticalized phrases with connecting function such as *the reason was* or *for this reason*), involving also relations with a noun phrase as its argument (Rysová and Rysová, 2015).

The first version of the annotation of discourse relations in the data of the Prague Dependency Treebank was published in 2012 as the Prague Discourse Treebank 1.0 (Poláková et al., 2012) and described in detail in Poláková et al. (2013). An updated version of the annotation of discourse relations of the same data was published in the Prague Dependency Treebank 3.0 (Bejček et al., 2013), with newly annotated second relations (see Example 1) and newly added rhematizers as parts of connectives (the updates were reported in Mírovský et al., 2014). A detailed study dedicated to different aspects of discourse relations and coherence in Czech, elaborating on various types of annotations of discourse-related phenomena in the data of the Prague Dependency Treebank, can be found in Zikánová et al. (2015). The most recent version of the annotated data, published as the Prague Discourse Treebank 2.0 (Rysová et al., 2016), newly brings annotation of discourse relations marked by secondary connectives. This last version of the annotations was used as the source data in the development of CzeDLex, as reported in the present paper. Table 1 summarizes the most significant changes of the annotation of discourse relations in various versions of the published data.

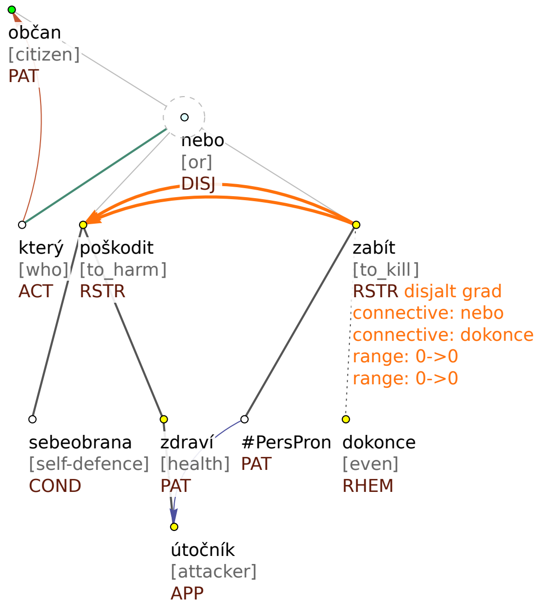


Figure 1: Annotation of discourse relations in PDiT 2.0. The relations are represented by two orange arrows connecting roots of the arguments. Information about the discourse types and connectives is given at the starting node of the relations.

Example 1 offers an illustration of discourse relations annotated in PDiT 2.0. It contains two intrasentential discourse relations – a disjunctive alternative expressed by the connective *nebo* [*or*], and a gradation expressed by the connective *dokonce* [*even*]; the tectogrammatical tree of the relevant part of the sentence, along with the discourse annotation, is depicted in Figure 1.

(1) *Občané, kteří v sebeobraně poškodili zdraví útočníka* ***nebo*** *ho* ***dokonce*** *zabili, bývají za své jednání často nespravedlivě stíháni.* (PDiT 2.0)

[*Lit.: Citizens who in self-defence harmed health of the attacker* ***or even*** *killed him, are often unfairly prosecuted for their actions.*]

# 3 Theoretical Issues

In this section, we first briefly compare our approach to the principles of development of related lexicons and then we provide a list of connective properties in CzeDLex, accompanied by description of necessary modifications made due to practical issues.

## 3.1 Inspiration from Other Lexicons

In the initial phase of the lexicon development, we kept in mind to be theoretically and technically as close to existing electronic lexicons of connectives as possible for the purposes of future lexicon linking and usability for translation. As stated earlier, the main source of inspiration was the German machine-readable Lexicon of Discourse Markers, DiMLex, developed since 1998 (Stede and Umbach, 1998) and continuously enhanced (DimLex 2, Scheffler and Stede, 2016). Like DiMLex, CzeDLex is encoded in XML (see Section 4.1 below), covers the part-of-speech, syntactic and semantic properties of the items described. Semantic properties are described via similar frameworks – a variant of the PDTB sense taxonomy (the PDTB version 3 for DiMLex versus Prague adjustments of the PDTB version 2.0 for CzeDLex). The core of the category of discourse connectives/markers is determined quite in agreement, although independently: DiMLex adopts the definition from Pasch et al. (2003), CzeDLex is inspired by the definition in the PDTB (see Section 2). Items covered in DiMLex include also several prepositions, or, more precisely, adpositions (*-halber, um ... Willen*), which is so far not the case for CzeDLex. In contrast, CzeDLex covers also some frequent secondary discourse connectives (similar to the "AltLex" category in the PDTB approach). Inclusion of both these groups of expressions in electronic inventories is quite a novel approach and can support further research on connectives in different languages and lexicographic projects. Nesting of lexicon entries in DiMLex follows the syntactic category of discourse markers. CzeDLex is structured differently, according to the discourse types (senses) of each lemma, see Section 3.2. The latter approach is also taken in the lexicon of French connectives, LexConn (Roze et al., 2012).

## 3.2 List of Connective Properties in CzeDLex

As PDiT 2.0 covers annotation both of primary and secondary connectives, CzeDLex contains both these groups. These two types of connectives differ lexico-syntactically as well as semantically and thus the linguistic information in the entries varies in several aspects. We first describe an entry of a primary connective and then for a secondary connective.

The theoretical basis of the structure of the lexicon entries and their properties has been adopted from the theoretical framework developed in Mírovský et al. (2016b). The entries in CzeDLex are structured according to a two-level nesting principle. On the first level, entries are nested according to the lemma of a connective. Apart from the lemma and its approximate English translation, the level-one entry contains the following linguistic information:

- type of the connective (primary vs. secondary),
- structure of the connective (whether the connective is single like *a* [*and*], *ale* [*but*] or complex like *i když* [*even though*]),
- variants of the connective (variants may be of a different kind, cf. stylistic variants like *tedy* [*so.neutral*] vs. *teda* [*so.informal*] or orthographic variants like *protože* vs. *proto, že*, both meaning [*because*] or inflection variants, e.g. the form *čehož* is the second case form of the connective with the first case form *což* [*which*]),
- connective usages – a list of level-two entries representing semantico-pragmatic relations the connective expresses and their properties,

- non-connective usages – another list of level-two entries, representing contexts where the lemma does not function as a discourse connective (e.g. "mum and dad").

Level-two nesting for primary connectives reflects the discourse-semantic types (condition, opposition etc.).[1] It is the lemma in combination with the discourse type, not the lemma alone, which allows for searching for the connective's counterparts in translation and lexicon linkage. Entries for the individual semantic types of a connective (called "usages" in the data structure) then contain the following pieces of information:

- semantic type of the discourse relation (condition, opposition etc.),
- gloss (an explanatory Czech synonym),
- English translation,
- part of speech of the connective,
- argument semantics (for asymmetric relations like reason–result, it is necessary to determine whether the argument syntactically associated with the connective expresses reason (e.g. *protože* [*because*]) or result (e.g. *proto* [*therefore*])),
- ordering , i.e. position of the argument syntactically associated with the connective in relation to the other (external) argument (e.g. Czech coordinating conjunctions, adverbs and particles are placed in the linearly second argument),
- integration, i.e. placement of a connective in an argument (e.g. Czech subordinating conjunctions are placed at the beginning of a clause),[2]
- list of the connective modifications (a modified connective contains an expression further specifying the relation, e.g. *hlavně protože* [*mainly because*]),
- list of complex connectives containing the given connective (a complex connective contains two or more connective words like *a proto* [*and therefore*]),
- examples from PDiT (i.e. a context for the given discourse relation) and their English translations,
- is_rare (set to '1' for rare usages),
- register (formal, neutral, informal).

An entry for a secondary connective contains several modifications. On level one of the lexicon structure, entries are nested according to the lemma of the core word for a secondary connective (core words are words such as *důvod* [*reason*] in *z tohoto důvodu* [*for this reason*], *to je důvod, proč* [*that is the reason why*] etc., or *podmínka* [*condition*] in *podmínkou bylo* [*the condition was*], *za těchto podmínek* [*under these conditions*]). A level-two entry then contains the following properties (we list here the additional properties assigned only to the secondary connectives).

- syntactic characteristics of the structure (e.g. *z tohoto důvodu* [*for this reason*] is a prepositional phrase),
- dependency scheme (general pattern) for each structure (e.g. *z tohoto důvodu* [*for this reason*] = "z ((anaph. Atr) důvod.2)", i.e. a preposition *z* [*for*] plus an anaphoric attribute and the word *důvod* [*reason*] in genitive),
- realizations of the dependency scheme (e.g. *z tohoto důvodu* [*for this reason*], *z daných důvodů* [*for the given reasons*], *z uvedených důvodů* [*for the stated reasons*]).

### 3.3 Unifying Changes

The theoretically pure data schema of the lexicon (described shortly above) was slightly modified in the implementation of the lexicon in several aspects, making it more suitable for practical use. The most important changes involved:

(i) On the second level of the lexicon structure, the secondary connectives are nested not only according to the discourse type they express, but also according to the syntactic structure of similar surface realizations of the connective. A purer solution would result in a three-level hierarchy for the secondary connectives. This more practical solution keeps the data structure almost identical for the primary and secondary connectives.

(ii) The part of speech of the secondary connectives (their core words) should be on the first level, as it cannot differ in various connective or non-connective usages. On the other hand, the part of speech of a primary connective word can be differ-

[1] Level-two nesting of non-connective usages is based on the part of speech of the lemma.

[2] The names of the elements ordering and integration are taken from DiMLex.

ent (at least for connective vs. non-connective usages), and therefore it has to be placed at the second level. For unification reasons, the part of speech was placed at the second level also for the secondary connectives.

The positive impact of these modifications becomes probably most evident in querying the lexicon, significantly simplifying queries concerning both the primary and secondary connectives (we mention a querying tool later in Section 4.1).

## 4 Practical Implementation

This section describes the practical implementation of the lexicon in the Prague Markup Language framework (PML, see below) and advantages this choice brings. Two short examples show in detail how the data format looks like, to demonstrate a relative ease of using the PML formalism and possibly encourage others to use it in their practical research. We also shortly describe technical steps in the process of extracting the lexicon from the Prague Discourse Treebank and mention a few post-processing steps needed to improve the quality of the final data, and connective properties that need to be inserted into the lexicon manually.

### 4.1 Prague Markup Language

The primary format used for the Prague Dependency Treebank since version 2.0 is called the Prague Markup Language (PML, Hana and Štěpánek, 2012).[3] It is an abstract XML-based format designed for annotation of linguistic corpora, especially treebanks. It is completely independent of a particular annotation schema and can capture simple linear annotations as well as annotations with one or more richly structured interconnected annotation layers, dependency or constituency trees. The PML format has since been used for many other treebanks, most importantly the Prague Discourse Treebank but also the Prague Czech-English Dependency Treebank (Hajič et al., 2012), all treebanks in the HamleDT project (Zeman et al., 2015), and many others.

Representing data in the PML format immediately brings the following advantages:[4]

- The data can be browsed and edited in TrEd, a fully customizable tree editor (Pajas and Štěpánek, 2008). TrEd is written in Perl and can be easily customized to a desired purpose by extensions that are included in the system as modules.[5]
- The data can be processed using scripts written in btred – a command line version of TrEd.
- The data can be searched in the PML-TQ (Prague Markup Language–Tree Query, Pajas and Štěpánek, 2009), a powerful, yet user friendly, graphically oriented system for querying linguistically annotated treebanks.

The listing in Figure 2 is a short example from the PML-schema for CzeDLex, i.e. from the definition of the format of the lexicon data in the PML, namely a definition of the format for level-one entries (the lemmas). Notice the declarations of roles (`role="#NODE"`, `role="#CHILDNODES"`, lines 2 and 9), defining which data structures should be understood (i.e. represented) as tree nodes, and also the declaration of the identifier role (`role="#ID"`, line 3), defining which element should be understood as the key for the records.

The following example shows the respective part of the resulting lexicon entry for the connective *tedy* [*therefore*]:

```
<lemma id="l-tedy" pdt_count="576">
    (a level-one entry)
  <text>tedy</text> (the lemma itself)
  <type>primary</type> (vs. secondary)
  <struct>single</struct>
        (vs. complex)
  <variants>
    <variant register="informal"
             pdt_count="1">
      teda (an informal variant)
    </variant>
  </variants>
  <usages>
    (lists of connective and
     non-connective usages)
  </usages>
</lemma>
```

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

[4] And, of course, as the PML format is technically an XML, any general XML tool can be used for the data as well.

[5] Such a module was used also for the annotation of discourse relations in the PDT, see Mírovský et al. (2010).

```
01 <type name="c-lemma.type">
02   <structure role="#NODE">
03     <member as_attribute="1" name="id" role="#ID" required="1">
         <cdata format="ID"/></member>
04     <member as_attribute="1" name="pdt_count">
         <cdata format="nonNegativeInteger"/></member>
05     <member name="text" required="1"><cdata format="any"/></member>
06     <member name="type" type="c-type.type"/>
07     <member name="struct" type="c-struct.type"/>
08     <member name="variants" type="c-variants.type"/>
09     <member name="usages" type="c-usages-all.type" role="#CHILDNODES"/>
10   </structure>
11 </type>
```

Figure 2: A small piece from the PML-schema for CzeDLex, defining the data structure for the level-one entry – a lemma.

Similar type definitions need to be provided for all other parts of the lexicon data structure, i.e. for the types referred to in Figure 2 (such as `type="c-variants.type"`, line 6) and all other data types needed in the lexicon.

Figure 3 shows the lexicon loaded in the tree editor TrEd, allowing an annotator to make manual changes in the data. It displays an entry for the lemma *tedy* [*so*], with an opened dialog window for editing the connective usage representing the discourse type reason–result, and a roll-down list of available options for the value of the element `arg_semantics`. Individual lemmas (level-one entries), lists of connective usages, lists of non-connective usages, and individual usages (level-two entries) are represented by tree nodes.

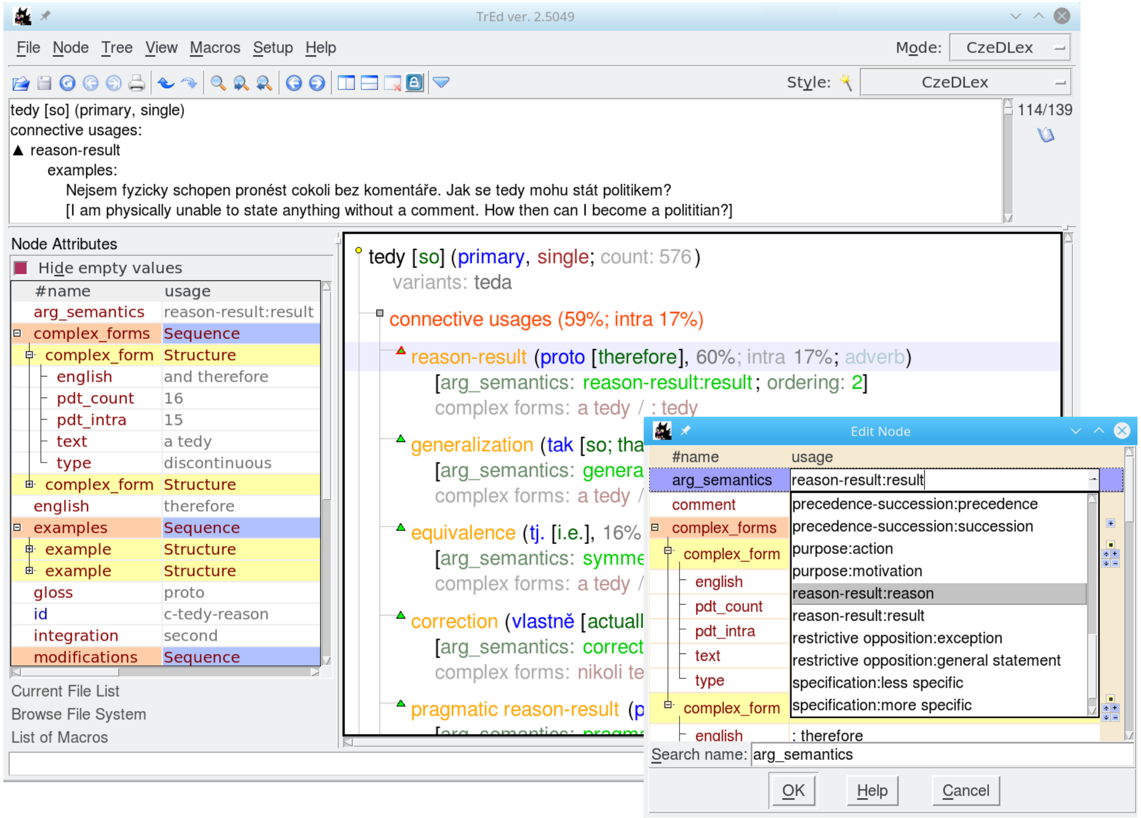
Using the PML for the lexicon CzeDLex brings, apart from the three advantages named above, another possibility – the lexicon can be easily interlinked with the source data, i.e. the Prague Discourse Treebank, by adding identifiers of the lexicon entries to the respective places in the treebank, using so called PML references. The query system PML-TQ then allows for incorporating information both from the treebank and the lexicon into a single query, allowing – for example – to search for all discourse relations in the treebank with connectives that have the ability to express (in different contexts) more than 2 different discourse types (senses).[6]

### 4.2 Data Extraction

The automatic extraction of the lexicon entries from the data of the Prague Discourse Treebank 2.0 (PDiT) was implemented in btred, a command line version of the tree editor TrEd. As an input, it used lists of lemmas accompanied by lists of variants, complex forms and modifications, which were created manually from the list of all connectives annotated in PDiT. In this all-connective list, each different string of words (e.g. *ale* [*but*] vs. *ale zároveň* [*but at the same time*] vs. *ale také* [*but also*]) formed a separate item. Primary and secondary connectives were distinguished automatically (in over 20 thousand annotated discourse relations in the treebank, there were approx. 700 different items for primary connectives and 350 for the secondary ones). Then, starting from the most frequent single connectives as lemmas, their variants, complex forms and modifications possibly belonging together under this lemma were selected manually from this all-connective list.

Based on this material, the script processed the whole data of PDiT, found all occurrences of the lemmas (and their variants etc.) and sorted them into the lexicon according to their type of usage (connective vs. non-connective) and the discourse type of the relations (or the part of speech for non-connective usages). For each usage, the part of speeach was automatically set and a number of the shortest examples were collected (the annotators later chose the most suitable ones and added their English translations). For each connective usage,

[6] See Mírovský et al. (2014) and Mírovský et al. (2016a) for examples of using the PML-TQ for searching in discourse-annotated treebanks (PDT 3.0 and PDTB 2.0, respectively).



tedy [so] (primary, single)
connective usages:
▲ reason-result
examples:
Nejsem fyzicky schopen pronést cokoli bez komentáře. Jak se tedy mohu stát politikem?
[I am physically unable to state anything without a comment. How then can I become a polititian?]

| #name | usage |
|---|---|
| arg_semantics | reason-result:result |
| ⊟ complex_forms | Sequence |
| ⊟ complex_form | Structure |
| english | and therefore |
| pdt_count | 16 |
| pdt_intra | 15 |
| text | a tedy |
| type | discontinuous |
| ⊞ complex_form | Structure |
| english | therefore |
| ⊟ examples | Sequence |
| ⊞ example | Structure |
| ⊞ example | Structure |
| gloss | proto |
| id | c-tedy-reason |
| integration | second |
| modifications | Sequence |

tedy [so] (primary, single; count: 576)
variants: teda
connective usages (59%; intra 17%)
reason-result (proto [therefore], 60%; intra 17%; adverb)
[arg_semantics: reason-result:result; ordering: 2]
complex forms: a tedy / : tedy
generalization (tak [so; tha
[arg_semantics: genera
complex forms: a tedy /
equivalence (tj. [i.e.], 16%
[arg_semantics: symme
complex forms: a tedy /
correction (vlastně [actuall
[arg_semantics: correct
complex forms: nikoli te
pragmatic reason-result (p

| #name | usage |
|---|---|
| arg_semantics | reason-result:result |
| comment | |
| ⊟ complex_forms | |
| ⊟ complex_form | |
| english | |
| pdt_count | |
| pdt_intra | |
| text | |
| type | |
| ⊟ complex_form | |
| english | : therefore |

precedence-succession:precedence
precedence-succession:succession
purpose:action
purpose:motivation
reason-result:reason
reason-result:result
restrictive opposition:exception
restrictive opposition:general statement
specification:less specific
specification:more specific

Figure 3: CzeDLex opened in the tree editor TrEd.

in moste cases, the argument semantics and ordering were assigned according to the orientation of the discourse arrow and position of the connective in an argument. Numbers of occurrences in PDiT were added to all individual variants, complex forms and modifications, as well as to connective and non-connective usages (level-two entries) and the whole lemmas (level-one entries).

After the lexicon was extracted from the annotated treebank, a few automatic or semi-automatic post-processing and data validity checking steps were performed. All counts of appearances of various lexicon data structures in the source treebank data have been checked (e.g. if counts of individual connectives sum up to counts of the usages and the lemmas). Another important verifying step checked for each complex form (e.g. *ale také* [*but also*]) that its basic lemma (the respective level-one entry, say *ale* [*but*]) appeared in the treebank with the same discourse type. If not, the complex form was removed from that lemma (being for the moment left as a complex form of the other lemma forming the complex form, in our case *také* [*also*]). If the complex form was by this process removed from all its basic lemmas, a new level-one entry for this complex form was created, with the value complex in the element struct.

Several properties required manual work, as the treebank data either did not contain this information at all (English translations, Czech synonyms, register, rareness, constituency-based syntactic characteristics of secondary connectives, structure) or the data were not big enough to cover all existing possibilities (integration, dependency scheme, sometimes ordering).

## 5 Conclusion

We presented the development process and implementation of an electronic lexicon of discourse connectives in Czech (CzeDLex). First, theoretical lexicographic aspects of building a lexicon for both primary and secondary connectives were addressed.

Second, the practical approach was discussed, starting with the description of the data format used – the Prague Markup Language – and advantages this choice brings. We followed by an elaboration on the actual process of exploiting the Prague Discourse Treebank 2.0 – a large corpus manually annotated with discourse relations – to build the raw basis of the lexicon, with subsequent automatic and manual checks, corrections and additions. To make the lexicon readable for non-Czech speakers, all names of elements, attributes and their values (with the obvious exception of Czech word entries and Czech corpus examples) are in English. In addition, each entry in Czech was supplemented by its English translation, including all corpus examples.

The first version of CzeDLex will be published this year in the Lindat/Clarin repository[7] under the Creative Commons license. It will cover an essential part of the connectives used in the Prague Discourse Treebank 2.0.[8] The second version of CzeDLex, planned to be published next year, will cover all connectives annotated in the treebank.

## Acknowledgment

The authors gratefully acknowledge support from the Ministry of Education, Youth and Sports of the Czech Republic (project COST-cz LD15052), and the Grant Agency of the Czech Republic (project GA17-06123S). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). The authors are also grateful for inspiration coming from meetings and work realized within the European project TextLink (COST Action IS1312).

## References

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague Dependency Treebank 3.0. Data/software.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of LREC 2012*. ELRA, European Language Resources Association, İstanbul, Turkey, pages 3153–3160.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, and Zdeňka Urešová. 2006. Prague Dependency Treebank 2.0. Data/software.

Jirka Hana and Jan Štěpánek. 2012. Prague Markup Language Framework. In *Proceedings of LAW 2012*. Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, pages 12–21.

Jiří Mírovský, Pavlína Jínová, and Lucie Poláková. 2014. Discourse Relations in the Prague Dependency Treebank 3.0. In Lamia Tounsi and Rafal Rak, editors, *Proceedings of Coling 2014 System Demonstrations*. Dublin City University (DCU).

Jiří Mírovský, Lucie Mladová, and Zdeněk Žabokrtský. 2010. Annotation Tool for Discourse in PDT. In Chu-Ren Huang and Dan Jurafsky, editors, *Proceedings of Coling 2010*. Chinese Information Processing Society of China, Tsinghua University Press, Beijing, China, volume 1, pages 9–12.

Jiří Mírovský, Lucie Poláková, and Jan Štěpánek. 2016a. Searching in the penn discourse treebank using the PML-tree query. In Nicoletta Calzolari et al., editor, *Proceedings of LREC 2016*. European Language Resources Association, Paris, France, pages 1762–1769.

Jiří Mírovský, Pavlína Synková, Magdaléna Rysová, and Lucie Poláková. 2016b. Designing CzeDLex – A Lexicon of Czech Discourse Connectives. In *Proceedings of PACLIC 2016*. Kyung Hee University.

[7] http://lindat.cz

[8] All those that will have undergone all checks and manual additions by that time.

Petr Pajas and Jan Štěpánek. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In Donia Scott and Hans Uszkoreit, editors, *Proceedings of Coling 2008*. The Coling 2008 Organizing Committee, Manchester, pages 673–680.

Petr Pajas and Jan Štěpánek. 2009. System for Querying Syntactically Annotated Corpora. In Gary Lee and Sabine Schulte im Walde, editors, *Proceedings of the ACL–IJCNLP 2009 Software Demonstrations*. Association for Computational Linguistics, Suntec, pages 33–36.

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

Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Eva Hajičová, Jiří Mírovský, Anna Nedoluzhko, Magdaléna Rysová, Veronika Pavlíková, Jana Zdeňková, Jiří Pergler, and Radek Ocelák. 2012. Prague Discourse Treebank 1.0. Data/software.

Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. 2013. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of IJCNLP 2013*. Asian Federation of Natural Language Processing, Nagoya, pages 91–99.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In Nicoletta Calzolari et al., editor, *Proceedings of LREC 2008*. European Language Resources Association, Marrakech, pages 2961–2968.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: a French lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique* (10).

Magdaléna Rysová and Kateřina Rysová. 2014. The Centre and Periphery of Discourse Connectives. In Wirote Aroonmanakun, Prachya Boonkwan, and Thepchai Supnithi, editors, *Proceedings of PACLIC 2014*. Chulalongkorn University.

Magdaléna Rysová and Kateřina Rysová. 2015. Secondary connectives in the Prague Dependency Treebank. In Eva Hajičová and Joakim Nivre, editors, *Proceedings of Depling 2015*. Uppsala University.

Magdaléna Rysová, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková, and Šárka Zikánová. 2016. Prague Discourse Treebank 2.0. Data/software.

Tatjana Scheffler and Manfred Stede. 2016. Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In *Proceedings of LREC 2016*. European Language Resources Association, Paris, France.

Manfred Stede. 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.

Manfred Stede and Carla Umbach. 1998. DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In *Proceedings of Coling 1998*. Association for Computational Linguistics, pages 1238–1242.

Daniel Zeman, David Mareček, Jan Mašek, Martin Popel, Loganathan Ramasamy, Rudolf Rosa, Jan Štěpánek, and Zdeněk Žabokrtský. 2015. HamleDT 3.0.

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. 2015. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia.