PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 109 OCTOBER 2017 61–91

CzeDLex – A Lexicon of Czech Discourse Connectives

Jiří Mírovský, Pavlína Synková, Magdaléna Rysová, Lucie Poláková

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

CzeDLex is a new electronic lexicon of Czech discourse connectives, planned for publication by the end of this year. Its data format and structure are based on a study of similar existing resources, and adjusted to comply with the Czech syntactic tradition and specifics and with the Prague approach to the annotation of semantic discourse relations in text.

In the article, we first put the lexicon in context of related resources and discuss theoretical aspects of building the lexicon – we present arguments for our choice of the data structure and for selecting features of the lexicon entries, while special attention is paid to a consistent and (as far as possible) uniform encoding 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 – apart from the lexical form of the connective – a discourse-semantic type (sense) expressed by the given connective, which enables us to deal with a broad formal variability of connectives and is convenient for interlinking CzeDLex with lexicons in other languages.

Second, we introduce the chosen technical solution based on the Prague Markup Language, which 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.

Third, we describe the process of getting data for the lexicon by exploiting a large corpus manually annotated with discourse relations – the Prague Discourse Treebank 2.0: we elaborate on the automatic extraction part, post-extraction checks and manual addition of supplementary linguistic information.

 Corresponding author: mirovsky@ufal.mff.cuni.cz
Cite as: Jiří Mírovský, Pavlína Synková, Magdaléna Rysová, Lucie Poláková. CzeDLex – A Lexicon of Czech Discourse Connectives. The Prague Bulletin of Mathematical Linguistics No. 109, 2017, pp. 61–91.
doi: 10.1515/pralin-2017-0039.

1. Introduction

In connection with rapid development of corpora annotated with discourse relations for different languages and in various frameworks (Carlson et al., 2003, Prasad et al., 2008 (English), Oza et al., 2009 (Hindi), Zeyrek et al., 2010, Zeyrek and Kurfalı, 2017 (Turkish), Al-Saif and Markert, 2010 (Arabic), Danlos et al., 2012 (French), Zhou and Xue, 2012, 2015 (Chinese), Stede and Neumann, 2014 (German), Iruskieta et al., 2013 (Basque), Da Cunha et al., 2011 (Spanish), to refer to just a few),[1] electronic lexicons of discourse connectives began to be built, although they are so far much less common. Electronic lexicons of discourse markers[2] are not only a useful tool in the theoretical research of text coherence/cohesion. Systematic information on discourse markers contributes to NLP tasks that involve processing of discourse relations (cf. e.g. Meyer et al., 2011, Stede, 2014 or Lin et al., 2014) and may help in machine translation, information extraction, text generation and other areas.

Our goal was to design and build an electronic lexicon of Czech discourse connectives, having in mind especially the following objectives:

- to contribute to the theoretical understanding of Czech connectives, and more generally, to understanding how text coherence/cohesion is established in Czech,
- to help in NLP tasks such as discourse processing, text generation and machine-translation, and
- to make the lexicon readable to a non-Czech speaker and linkable to existing lexicons of connectives in other languages.

There are several options how to actually build such a lexicon, i.e. how to fill it with data, from consulting existing printed lexicons, to using translation from lexicons in other languages, to exploiting existing discourse-annotated corpora in the given language. We have chosen the last option, as a large discourse-annotated treebank – the Prague Discourse Treebank 2.0 (Rysová et al., 2016) – is available for Czech.

The present article summarizes, updates and extends information on the design and build-up of the lexicon of Czech discourse connectives – CzeDLex – that was previously given in Mírovský et al. (2016b) and Synková et al. (2017, in print).

The subsequent text is organized as follows: Section 2 gives an overview of related research and existing lexicons of discourse connectives and compares main properties of CzeDLex and the other resources. In Section 3, the Prague Discourse Treebank 2.0 is introduced. Section 4 specifies basic terms such as "connective" and describes the lexicon structure from the theoretical point of view, providing reasons for decisions

[1] See also a list of discourse annotated corpora compiled within the COST TextLink project: `http://www.textlink.ii.metu.edu.tr/corpus-view`.

[2] We use "discourse markers" as a broader term for expressions structuring discourse, and "discourse connectives" (DCs) as a narrower term for expressions signalling semantico-pragmatic relations between two abstract objects (see 4.1).

behind the lexicon design. Section 5 describes the data format and application framework selected for the implementation of the lexicon, and presents also the process of extracting the lexicon from the data of the Prague Discourse Treebank 2.0, including subsequent checks, manual corrections and additions.

2. Related Research, Other Lexicons of Connectives

In this section, we put CzeDLex in context of current lexicography and compare it to other existing lexicons of connectives or expressions to a certain extent overlapping with some types of connectives.

Generally, lexicons (or dictionaries) may be of various kinds, reflecting different linguistic aspects. Traditionally, lexicons are characterized according to the number of languages they involve (monolingual, bilingual, multilingual dictionaries), their coverage (a general dictionary, a dialect dictionary, a sociolect dictionary reflecting e.g. colloquial language, adolescent language etc.), aspects of linguistic structure (an orthographic dictionary, a pronunciation dictionary, a frequency dictionary, a phraseological dictionary), the segment of the vocabulary (a dictionary of neologisms or a loan-word dictionary) or the group of users (a language learner's dictionary). For more details, see Hausmann (1985).

In this respect, lexicons of discourse markers/connectives represent a part of a specific lexicographic domain: in contrast to the majority of dictionaries/lexicons, they describe the synsemantic part of vocabulary (i.e. grammatical words, function words). As such, these lexicons are in fact lists of possible forms that can express one certain function in a language. These functional lexicons are so far much rarer and even more so in the Czech context. For other languages, there are similarly targeted lexicons, let us mention e.g. German lexicographic projects: Lexikon deutscher Konjuktionen (Buscha, 1989), Lexikon deutscher Partikeln (Helbig, 1988), Präpositionen (Schröder, 1986), Modalwörter (Helbig and Helbig, 1990) etc. Regarding the connective/discourse marker category, the printed resources include Dictionary of link words in English discourse (Ball, 1993), or the German two-volume Handbuch der deutschen Konnektoren (HdK, Pasch et al., 2003; Breindl et al., 2015).

Another specificity of the last years is the machine-readable form of such functional resources and the intention (and often the primary goal) to use these resources in various NLP tasks. Apart from the "standard" digitalized monolingual or translation dictionaries for a large scope of users, there are, mostly corpus-based, electronic projects assembling vocabulary with a specific function (e.g. evaluative language in the Czech SubLex, Veselovská and Bojar, 2013), or mining morphosyntactic annotation, e.g. valency properties of verbs (CzEngVallex, Urešová et al., 2016, for Czech and English) and similar.

CzeDLex may thus be described as an electronic corpus-based resource of Czech discourse connectives, containing English connective equivalents, reflecting written journalistic Czech language of the PDiT 2.0 texts (see Section 3) that provides func-

tional descriptions of the expressions and phrases it covers. This includes morphosyntactic information, usage and meanings of the connectives in particular contexts and their frequencies in the underlying dataset (for more details, see Section 4.3).

Such a placement in the general typology of lexicographic projects puts CzeDLex right next to other newly emerging electronic lexicons of discourse markers or connectives. As far as we know, there are nowadays only few such projects, but the field is quickly growing and new projects arise every year now.[3] Perhaps one of the oldest electronic lexicons of discourse markers was the first version of DiMLex for German (Stede and Umbach, 1998), further, there is LexConn for French (Roze et al., 2012), DPDE for Spanish,[4] LICO for Italian (Feltracco et al., 2016) and others.[5] As some aspects of these lexicons served as a source of inspiration for the development of CzeDLex, we describe these lexicons and especially DiMLex in more detail later in this section.

From the Czech lexicographic projects, CzeDLex can be partly compared to the work of F. Čermák (2007, 2009). Secondary connectives in CzeDLex (i.e. expressions such as *z tohoto důvodu* [*for this reason*], for details see Section 4.1.1 below) to some extent overlap with phrases and idioms that are elaborated for Czech in his lexicon of Czech phrases and idioms. It consists of four volumes, dealing with 1. Comparisons, 2. Non-verbal expressions, 3. Verbal expressions and 4. Sentential expressions (see Čermák, 2009). Secondary connectives and Čermák's phrases and idioms presented in the lexicons overlap only slightly, but it is interesting to look at how these expressions are treated in various approaches.

The lexicon of phrases and idioms in Czech contains full and reduced lexicon entries. The full ones are for frequent expressions and the reduced ones for expressions with a lower frequency. The full entries contain various types of linguistic information such as stylistic characteristics, grammatical characteristics, intonation, context, valency, explanation of meaning, exemplification, synonyms or foreign language equivalents. The choice of entries (sorted in the alphabetical order) is based on corpus data (which is the same for CzeDLex). In this way, the lexicon aims to describe the current situation in the field of phraseology.

As an example of a sentential phrase in Čermák's lexicon, we find e.g. the phrase *Mám k tomu své/svý důvody.* [*lit.: I have my reasons for this.*] (which in PDiT 2.0 functions as a connective and was therefore included into CzeDLex under the connective phrases containing the word *důvod* [*reason*]). For a given phrase, the lexicon of phrases and idioms provides an explanation of its meaning, a context in which the phrase may

[3] To support building of such inventories of connectives in different European languages and to devise ways of interlinking their entries is one of the goals of the COST TextLink project, see `http://textlink.ii.metu.edu.tr`.

[4] `http://www.dpde.es`

[5] Compare also a list of inventories of discourse-structuring devices at `http://www.textlink.ii.metu.edu.tr/dsd-view`.

be used, and a synonymous construction. CzeDLex approaches similar phrases from a different perspective, namely in terms of coherence (i.e. we focus on the function the phrase has for the text coherence). Therefore, in CzeDLex, we deal with semantic discourse types expressed by the phrase, its Czech synonyms and English equivalent/s.

In the rest of this section, we compare the most important properties of some other existing connective lexicons to the properties of CzeDLex. During the design process of CzeDLex, the points of departure of similar projects were particularly important because of future lexicon interlinking and their usability for translation. We therefore aimed to be theoretically and technically as close to existing electronic lexicons of connectives as possible. As mentioned earlier, the main source of inspiration was the German machine-readable Lexicon of Discourse Markers, DiMLex (Stede and Umbach, 1998), developed in Potsdam and continuously enhanced (DiMLex 2, Scheffler and Stede, 2016). CzeDLex and DiMLex are indeed closely related in several basic aspects:

- they are both encoded in an XML-based format,
- the core of the delimitation of the category of discourse connectives/discourse markers is very similar,
- both cover part-of-speech, syntactic and semantic properties of the items they describe,
- semantic properties of the connectives are described via highly compatible frameworks – the sense taxonomy used in the Penn Discourse Treebank (Prasad et al., 2008) vs. its close Prague variant,
- both reflect ambiguity issues and record also non-connective usages.

On the other hand, different development processes of these inventories and different grammatical tradition (mostly in morphology) in discourse marker description resulted in several discrepancies between the two projects: Regarding the development process, DiMLex is being developed since 1998 and it is largely inspired by the extensive research project Handbuch der Deutschen Konnektoren (HdK; Pasch et al., 2003). CzeDLex is based upon the Prague Discourse Treebank 2.0, its annotation of discourse relations, syntactic analysis and part-of-speech tagging principles. The definition of a connective in DiMLex adopts five criteria from the HdK, M1-M5,[6] but drops the M2 criterion, as several (cca 25) prepositions, or, more precisely, adpositions (also postpositions, e.g. *–halber* and "Zirkumpositionen", e.g. *um ... Willen*), were considered discourse connectives and added to the lexicon. The CzeDLex connective definition is based on the Penn Discourse Treebank (PDTB) definition as a predicate of a binary

[6] (M1) X cannot be inflected. (M2) X does not assign case features to its syntactic environment. (M3) The meaning of X is a two-place relation. (M4) The arguments of the relation (the meaning of X) are propositional structures. (M5) The expressions of the arguments of the relation can be sentential structures (Scheffler and Stede, 2016).

relation opening positions for two text spans as its arguments and signalling a semantic or pragmatic relation between them (see 4.1 for details or compare Mírovský et al., 2016b). Prepositions are so far not included, but CzeDLex covers also some frequent secondary connectives (similar to the "AltLex" category in the PDTB approach). Some earlier work on more complex connective expressions with referential components in Czech can be found in Poláková et al. (2012) and mainly in Rysová and Rysová (2015), and for German, a pilot study of the anaphoric connective *demzufolge* [best translated as *accordingly, as a result, consequently*] is given in Stede and Grishina (2016). DiMLex now contains 275 German connectives of current use and the authors claim that the coverage is complete.

Nesting of lexicon entries in DiMLex follows the syntactic category of discourse markers. In this aspect, lemmas of connectives in CzeDLex are structured differently, according to discourse types (senses) they convey. The latter approach is also taken in the French LexConn (Roze et al., 2012), cf. e.g. several entries for the expression *alors*.

Semantic properties of the connectives are described via very similar frameworks: a variant of the PDTB sense taxonomy – PDTB 3.0 – for DiMLex versus Prague adjustments of the PDTB version 2.0 (see Table 1) for CzeDLex. In addition, DiMLex 2.0 was recently enriched by semantic relations according to more discourse frameworks, it lists all possible semantic/pragmatic characteristics of a given connective token also according to the frameworks of the Rhetorical Structure Theory (RST; Mann and Thompson, 1988b) and the Segmented Discourse Representation Theory (SDRT; Asher, 1993), and the grammar book of Helbig and Buscha (1984).

3. Prague Discourse Treebank 2.0

The Prague Discourse Treebank 2.0 (Rysová et al., 2016) 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 manual multilayer annotation of approx. 50 thousand sentences of Czech journalistic texts from 1990's. The Prague Dependency Treebank contains 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, pronominal and nominal 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 (for a definition of connective, see 4.1).

The set of discourse types (see the complete list in Table 1) is inspired by the Penn Discourse Treebank 2.0 sense hierarchy (Prasad et al., 2008) and the syntactico-semantic labels used for representation of compound sentences on the tectogrammatical layer.

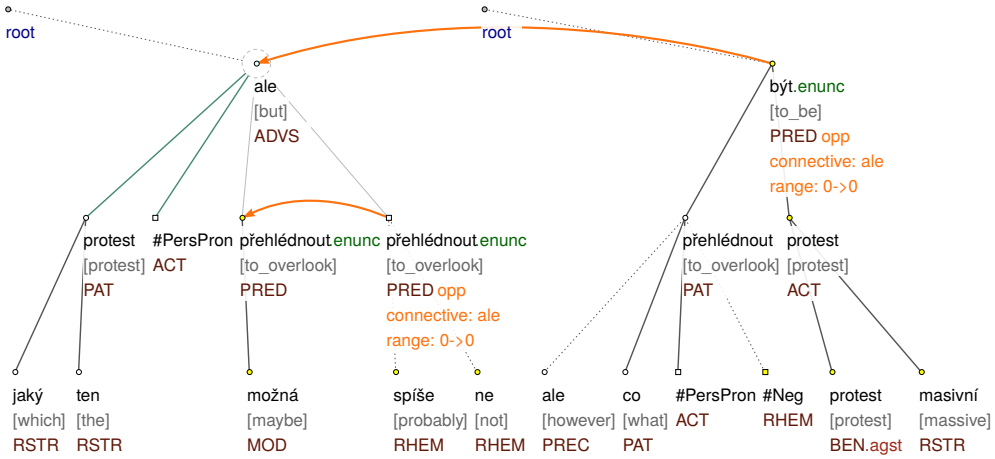

Figure 1. Example of an intra-sentential and an inter-sentential discourse relation in PDiT 2.0. Both relations are represented by thick curved arrows connecting roots of the arguments. Information about the semantic discourse types, connectives and range of the arguments is given at the starting nodes of the relations.

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 (PDiT 1.0; Poláková et al., 2012b) and described in detail in Poláková et al. (2013).

(1) *Možná jsem nějaký ten protest přehlédl,* ***ale*** *spíše ne. Co jsem* ***ale*** *přehlédnout nemohl, byly masivní protesty proti protestům.* (PDiT 2.0)
[*Lit.: Maybe I overlooked some of the protests* ***but*** *probably not. What I* ***however*** *could not overlook, were massive protests against protests.*]

An updated version of the annotation of discourse relations of the same data was published in the Prague Dependency Treebank 3.0 (PDT 3.0; Bejček et al., 2013), with newly annotated second relations[7] and more systematic annotation of focusing particles (such as *also*, *too*) as parts of connectives of *conjunction* relation. A new attribute *discourse_special* was introduced to capture three special roles of phrases: headings

[7] Note that – unlike in the Penn Discourse Treebank approach – second relations annotated in the Prague Dependency Treebank 3.0 and in the Prague Discourse Treebank 2.0 only involve cases where different relations (in the term of semantic discourse type) between the same arguments are explicitly expressed by two different connectives (e.g. relations *opposition* and *asynchronous* expressed by connectives *but* and *then*, respectively, in the sentence *He wanted to go there* ***but then*** *he changed his mind.*). Second relations as they are understood in the Penn Discourse Treebank approach – i.e. two relations expressed by a single connective – are not annotated in our data.

CONTRAST	EXPANSION
confrontation	*conjunction*
opposition	*conjunctive alternative*
restrictive opposition	*disjunctive alternative*
pragmatic contrast	*instantiation*
concession	*specification*
correction	*equivalence*
gradation	*generalization*
CONTINGENCY	**TEMPORAL**
reason–result	*synchrony*
pragmatic reason–result	*precedence–succession*
explication	
condition	
pragmatic condition	
purpose	

Table 1. Semantic types of discourse relations in PDiT 2.0 and CzeDLex

(replaced the attribute *is_heading* from PDiT 1.0), metatext (text not belonging to the original newspaper text, produced during the creation of the corpus), and captions of pictures, graphs etc. (the updates were reported in Mírovský et al., 2014). Genres of documents were also annotated in the PDT 3.0 (and reported in Poláková 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).

Annotations published in PDiT 1.0 and in the PDT 3.0 involved explicit discourse relations expressed by connectives belonging mostly to conjunctions, adverbs, particles and punctuation marks, some of them were formed also by multi-word phrases.[8] In 2014, discourse connectives were divided into primary and secondary according to their degree of grammaticalization (Rysová and Rysová, 2014, 2015), see 4.1.1 below.

[8] A detailed list of expressions involved in the PDiT 1.0 and PDT 3.0 annotations: (i) coordinating conjunctions: e.g. *a* [*and*], *ale* [*but*], *však* [*but*], (ii) subordinating conjunctions: e.g. *ačkoliv* [*although*], *protože* [*because*], (iii) particle expressions (including rhematizers): e.g. *ovšem* [*however*], *zkrátka* [*shortly*], (iv) adverbs: e.g. *potom* [*then*], *stejně* [*equally*], (v) some prepositions with demonstrative pronouns: e.g. *kromě toho* [*except for this*], *k tomu* [*in addition to this*], *tím* [*by this*], (vi) some types of idiomatic multiple-word connective means formed by linking of different expressions: e.g. *na jedné straně* [*on the one hand*], *stručně řečeno* [*in short*], *jinými slovy* [*in other words*], (vii) elements formed by letters or numbers expressing enumeration: e.g. a), b), 1., 2.; (viii) two punctuation marks: colon and dash (see Poláková et al., 2012a). These connectives are described in detail in Poláková (2015).

	PDiT 1.0 (2012)	PDT 3.0 (2013)	PDiT 2.0 (2016)
Primary connectives[9]	yes	updated	updated
Headings	yes	yes	yes
Second relations		yes	updated
Focusing particles		yes	yes
Captions, metatext		yes	yes
Genres of documents		yes	yes
Secondary connectives			yes

Table 2. Principal changes in the annotation of discourse relations and related phenomena in various published versions of the data. Each new version also brought fixes of annotation errors.

This new division is reflected in the newest published version of the Prague discourse annotation – the Prague Discourse Treebank 2.0 (PDiT 2.0; Rysová et al., 2016). Specifically, PDiT 2.0 contains a minor revision of the previous annotation (some types of connectives such as *kromě toho* [*except for this*] were re-annotated as secondary connectives) and annotation of discourse relations expressed by a new set of secondary connectives was added.

Table 2 summarizes the most significant changes of the annotation of discourse relations in the various versions of the published data. The last version – the Prague Discourse Treebank 2.0 – was used as the source data in the development of CzeDLex, as reported in the present article.

4. Theoretical Aspects

4.1. A Connective

One of the basic decisions in building a lexicon of discourse connectives concerns the delimitation of the connective category. In accordance with the Prague tradition of discourse annotation and the approach used for the annotation of PDiT 2.0, we understand a discourse connective 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.[10]

[9] We use the term "primary connectives" here in a simplified way, as this term was first used and defined in 2014. However, the annotations of explicit connectives in PDiT 1.0 (Poláková et al., 2012b) and the PDT 3.0 (Bejček et al., 2013) roughly correspond to this class of expressions; see footnote 8 for a detailed list.

[10] A similar approach was used in the PDTB, cf. Prasad et al. (2008).

The two connected text segments are defined according to Asher (1993) as abstract objects expressing events, states, situations, etc. Syntactically, abstract objects (discourse arguments) can be represented by various structures ranging from whole sentences or their combination, to simple clauses, to participial and infinitive constructions and nominal phrases. In PDiT 2.0, annotation of discourse relations was syntactically restricted to verbal arguments (i.e. whose basis is a finite verb).[11] CzeDLex therefore includes connectives in relations with verbal arguments only.

4.1.1. Primary and secondary connectives

Discourse connectives in PDiT 2.0 are divided into primary and secondary ones, according to Rysová and Rysová (2014), as already mentioned in Section 3. Primary connectives were defined as grammaticalized expressions such as *because* or *therefore* whereas secondary connectives were established as not (yet) fully grammaticalized structures with connecting function such as *except for this*, *the reason was* or *for this reason*.

CzeDLex contains both types of connectives. They, however, differ in many important aspects that need to be reflected in the lexicon design: lemmatization, syntactic characteristics, part-of-speech appurtenance, position of the arguments and argument integration (i.e. the position of a connective in the argument). Many secondary connectives may be inflected (*for this reason – for these reasons*; *the condition is – the conditions were* etc.) and they exhibit – at least in Czech – a high degree of variation (*důvod je* vs. *důvodem je* [*the reason is*: nominative vs. instrumental], both variants in Czech are equivalent).

4.1.2. Complex forms and modified connectives

Discourse connectives often occur in complex and/or modified forms (see Rysová, 2015). Complex forms consist of two or more connective words (i.e. words that can be connectives by themselves) that all participate on expressing the given discourse meaning (semantic discourse type, sense). Complex forms occur either in a single argument (*a proto* [*and therefore*]) or they may form correlative pairs (*buď nebo* [*either or*]).

Modified connectives contain an expression of an evaluative, modal or intensifying nature that further specifies/modifies the discourse relation, without changing its semantic type (*hlavně protože* [*mainly because*] or *možným důvodem je* [*a possible reason is*]).

[11] The annotation of secondary connectives in PDiT 2.0 took into account also arguments formed by noun phrases. These cases were annotated as notes at the core words of the secondary connectives, without the full annotation of the discourse relations (the whole connective, the arguments and their extent are not marked); these cases are not included in CzeDLex.

Both complex and modified connectives are included in CzeDLex, as parts of entries for the respective single connectives (for details and exceptions, see 4.3 below).

4.1.3. Non-connective usages

Most connective expressions (or, in case of secondary connectives, certain parts of them) exhibit a functional homonymy with expressions that have different functions in the text. Non-connective usages of these homonymous expressions can be categorized into several groups with specific properties:

- Expressions connecting mere entities (e.g. *towns **and** villages*) are not considered discourse connectives since they do not connect abstract objects (Asher, 1993).
- Expressions in the function of expressive, modifying or answer particles do not connect two abstract objects either, although their function belongs to the wider class of discourse markers in some contexts (e.g. ***So**, will you visit her? **Of course**.*).
- Homonyms of primary connectives sometimes function only as sentence constituents (mostly in the rhematic part of a sentence) and not as connectives (e.g. *Musíš to udělat úplně **jinak**.* [*lit.: You have to do it completely **otherwise**.*]).
  In contrast to the primary ones, the secondary connectives (or their parts) are always sentence constituents at the same time. However, their "core" words may also have a non-connective usage – cf. *The suggestion was rejected for procedural **reasons**.*

For each lexicon entry in CzeDLex, in addition to the list of connective usages, non-connective usages of the expression/phrase are listed at level two of the lexicon structure (see 4.2), along with their syntactic characteristics.[12]

4.2. Nesting of Lexicon Entries

The most important property of a discourse connective is its lexical form, and naturally the connectives in the lexicon are nested[13] on the first level (level one) according to their lemmas (which need to be representatively chosen for complex or modified connectives, and especially for secondary connectives, see below in 4.3).

Since we are building a lexicon of *discourse* connectives, the second most important property of a connective is the semantic discourse type the connective can convey (more precisely, a list of the discourse types). Therefore, on the second level (level two) of the lexicon structure, the entries are nested according to these semantic dis-

[12] A detailed analysis of "the degree of connectivity" of frequent Czech connectives according to the PDT 3.0 annotation can be found in Zikánová et al. (2015, pp. 161–162).

[13] By "nested" we mean organized, divided into individual entries.

course types.[14] This approach is justified also by a practical consideration: one of the primary uses of a lexicon of discourse connectives is machine translation. Interconnected lexicons of discourse connectives in different languages may help choose a correct translation of a connective in the given context (see e.g. Meyer and Poláková, 2013). The following observations suggest an answer to the question "which items (parts of records in the lexicons) should get connected?".

For translating a discourse connective to another language, it is not sufficient to only know the connective itself; for example, if we look up a translation of the English connective *while* into Czech in a publicly available online translation dictionary,[15] we get the following list:

- *zatímco*; *když* (synchronous events)
- *když*; *během toho, co* (synchronous events)
- *zatímco*; *kdežto*; *ale* (adversative relation [but])
- *i když*; *ačkoli*; *přestože* (concession [although])
- *(nějaký) čas*; *chvíle*; *chvilka* (noun)

...which is an ambiguous result and – as we can see – most of the options differ in the semantics of the connectives, which is very close to the discourse semantic type.[16]

On the other hand, to correctly translate a connective in context, it is not sufficient to know only the semantic discourse type the connective conveys either: if we try to "translate" the connective *while* based only on the fact that it is – in the given case – expressing e.g. the sense of *Contrast* in the PDTB taxonomy, and if we assume that in the Prague taxonomy the respective discourse type is *opposition*, we will find out that in PDiT 2.0, the relation of *opposition* is realized by 103 different connectives[17] – the most frequent of them are listed in Table 3.

We can conclude that to select a proper translation of a discourse connective, we need both the lexical information (the connective itself) and the semantic discourse type conveyed by the connective in the given context. This supports the chosen approach of nesting entries in the lexicon according to lexical forms of the connectives (level one) and semantic discourse types they convey (level two). These level-two entries are then to be mapped to their counterparts in other lexicons.[18]

---

[14] Non-connective usages of the connective words are nested according to their part of speech.

[15] `https://slovnik.seznam.cz`

[16] The part of speech of the connectives would not be of much help here – only one option ("noun") would be ruled out. For most other connectives, the part of speech would not help at all.

[17] including variants, complex forms and modifications

[18] There are of course many remaining issues. The linking is still not 1:1, lexicons use different definitions of "connectives", different taxonomies of semantic discourse types, different lists of features for entries in the lexicons, etc.

connective	count	connective	count
však	1 104	*nicméně*	36
ale	955	*sice ... však*	35
ovšem	197	*přitom*	32
sice ... ale	122	*aniž*	21
jenže	44	*a*	16
avšak	41	...	

Table 3. Most frequent connectives in PDiT 2.0 expressing the relation of opposition.

4.3. Connective Properties in CzeDLex

Based on the above considerations, the entries in CzeDLex are nested according to a two-level principle. We describe in detail properties of entries on these two levels here in 4.3.1 and 4.3.2.

4.3.1. Level-one

The level-one entry in the lexicon structure is represented by the lemma of the connective. Whereas selecting a representative lemma for primary connectives is usually a straightforward decision (see 4.3.2 for details about complex connectives), a suitable solution needs to be carefully thought of for secondary connectives.

There are, for example, many secondary connectives containing the word *reason* (*for this reason*, *that is the reason why*, *the reason is* etc.). We can consider the word *reason* their common "core" word, i.e. the word that most strongly signals the relation that the whole secondary connective expresses. In the lexicon structure, we group secondary connectives under lemmas of these "core" words, which are mainly nouns (*reason*, *condition*, *conclusion* etc.), secondary prepositions (*due to*, *because of*, *thanks to* etc.) and verbs (*to precede*, *to conclude*, *to sum up* etc.)

The first level entry as a whole is encoded in the element[19] ***lemma*** and contains the following information:

- element ***text***: the lemma of the connective
- element ***english***: an approximate English translation for a basic orientation; more precise translations are given in connection with semantic discourse types at level-two entries
- element ***type***: the type of the connective: *primary* vs. *secondary* (see 4.1.1)

[19] Some properties of the lexicon entries are encoded as XML elements, others as their attributes (see Section 5).

- element ***struct***: the structure of the connective: it signals whether the connective is *single* such as *proto* [*therefore*] or *complex* such as *jednak jednak* [*on the one hand on the other hand*]. The complex connectives are further differentiated in the attribute ***type***[20] according to their placement in the argument(s): complex connectives with parts occurring in both arguments (e.g. *jednak jednak* [*on the one hand on the other hand*] or *buď nebo* [*either or*]) are labeled *correlative*, while complex connectives with all parts occurring in a single argument are labeled *continuous* if no word can be inserted between the parts of the connective (e.g. the connective *i když* [*even if, although*]), or *discontinuous* if other words can occur between the connective parts (e.g. *a potom* [*and then*]).
- element ***variants***: a list of variants of the connective: they are further specified in the attribute ***type*** as *stylistic* (cf. neutral *tedy* [*so.neutral*] vs. informal *teda* [*so.informal*]) or *orthographic* (e.g. *mimoto* vs. *mimo to* [*both meaning: besides*]), or *inflection* (e.g. the form *čímž* [*by which*] is the instrumental form of the connective with the nominative form *což* [*which*])
- element ***conn-usages***: a list of connective usages – level-two entries
- element ***non-conn-usages***: a list of non-connective usages – level-two entries
- attribute ***id***: a lexicon-wide unique identifier of this level-one lexicon entry

4.3.2. Level two

For each level-one entry in the lexicon structure, its connective and non-connective usages are represented as level-two entries. In connective-usages, the discourse type is used as the base for nesting (reasons for this decision were given in 4.2), while in non-connective-usages (see 4.1.3), the part-of-speech appurtenance of the expressions is used.

If this rule were followed strictly, the depth of the lexicon structure for secondary connectives would increase to three levels, as these connectives often form different syntactic structures conveying the same discourse type that cannot be treated in a single unit – for example, both secondary connectives *for the following reason* and *that is the reason why* express the same semantic discourse type (*reason–result*) but differ in the argument semantics, i.e. the former signals the reason, while the latter signals the result (see the element *arg_semantics* below).

To keep the data structure identical both for primary and secondary connectives,[21] we keep the two-level structure also for the secondary connectives; they are therefore nested not only according to the discourse type they express, but also to their representative dependency scheme. This scheme is a general pattern for the connective structure – e.g. the secondary connectives *z tohoto důvodu* [*for this reason*], *z uvedených*

[20] It is an attribute *type* of the element *struct*, different from the element *type* above.

[21] which, for example, simplifies searching in the lexicon in the PML-Tree Query system (see Section 5)

důvodů [*for the given reasons*] or *z těchto důvodů* [*for those reasons*] are represented by the dependency scheme "z ((anaph. Atr) důvod.2)", i.e. a preposition *z* [*for*] plus an anaphoric attribute and the word *důvod* [*reason*] in genitive.

The second level entry of the lexicon is encoded in the element ***usage*** and contains the following information:

- element ***sense***: the discourse type (see possible values in Table 1)
- element ***scheme***: the dependency scheme (used for secondary connectives only)
- element ***gloss***: a Czech expression disambiguating the meaning of the connective (a synonym or an explanatory phrase)
- element ***english***: an English translation (the gloss in English)
- element ***pos***: the part-of-speech appurtenance of the connective (the lemma) in the given usage. Conjunctions are further distinguished in the attribute ***subpos*** as *coordinating* or *subordinating*.
- element ***syntax***: for secondary connectives, the part-of-speech characteristics of the core word is accompanied by a syntactic characteristics for the whole secondary connective represented by this usage (*nominal phrase*, *adjectival phrase*, *pronominal phrase*, *clause*, *adverbial phrase*, or *prepositional phrase*).
- element ***arg_semantics***: this characteristics specifies the semantics of the argument the connective occurs in. From the semantic perspective, there is a basic difference between symmetric and asymmetric discourse relations. While both arguments of a symmetric relation (i.e. *conjunction* or *synchrony*) share the same general semantic characteristics, asymmetric discourse relations (e.g. *reason–result* or *gradation*) hold between arguments that have different semantic nature (e.g. one argument expresses the reason, the other the result).[22] A connective of an asymmetric relation is characterized by its placement in one specific part of the relation it signals. For example, the coordinating conjunction *tedy* [*thus*] signals the result, while *totiž* [*because*] signals the reason. Similarly, the subordinating conjunctions *než* [*until*] and *když* [*when*] can be used for signalling *precedence–succession* – the former occurs in the argument expressing the event happening later, while the latter occurs in the argument expressing the earlier event. Table 4 gives an overview of all possible values for the attribute *arg_semantics*. For sym-

[22] In some approaches, the discourse types of the relations are different (e.g. Sanders et al. (1992) distinguish *Cause-Consequence* and *Consequence-Cause*, the PDTB 2.0 (Prasad et al., 2007) differentiates *Cause:reason* and *Cause:result* according to the argument order), in other approaches the relation remains the same, but some conventions marking ordering of the reason and the result are applied (e.g. in the Prague approach, there is only one *reason–result* relation, but the reason part of the relation is indicated by the starting point of the arrow (cf. Zikánová et al., 2015); the ISO standard (Prasad and Bunt, 2015) introduces only one *Cause* relation as well, the asymmetry of the relation is represented by specifying argument semantics in the definition of the relation). In the Rhetorical Structure Theory (Mann and Thompson, 1988a), the difference in the (a)symmetry of relations is captured by the feature of nuclearity (symmetric relations are multinuclear, while asymmetric ones have a nucleus and a satellite).

relation	argument semantics
concession	*concession:expectation* *concession:contra-expectation*
condition	*condition:condition* *condition:result of condition*
correction	*correction:claim* *correction:correction*
explication	*explication:claim* *explication:argument*
generalization	*generalization:more specific* *generalization:less specific*
gradation	*gradation:lower degree* *gradation:higher degree*
instantiation	*instantiation:general statement* *instantiation:example*
pragmatic condition	*pragmatic condition:pragmatic condition* *pragmatic condition:result of pragmatic condition*
pragmatic reason-result	*pragmatic reason-result:pragmatic reason* *pragmatic reason-result:pragmatic result*
precedence-succession	*precedence-succession:precedence* *precedence-succession:succession*
purpose	*purpose:action* *purpose:motivation*
reason-result	*reason-result:reason* *reason-result:result*
restrictive opposition	*restrictive opposition:general statement* *restrictive opposition:exception*
specification	*specification:less specific* *specification:more specific*
all other relations	*symmetric*

Table 4. Possible values of the argument semantics (attribute arg_semantics).

metric relations, the element *arg_semantics* has the value *symmetric*. For complex correlative connectives forming level-one entries, the value is given for the second part of the connective.

- element ***ordering***: signals the linear order of the argument the connective occurs in (relatively to the other – external – argument).[23] In the majority of cases, ordering is connected with the part-of-speech characteristics – coordinating conjunctions, adverbs and particles are placed in the second argument in the linear order, while subordinating conjunctions can be placed in either of the arguments. There are, however, exceptions – e.g. the particle *nejenže* [*not only that*] which occurs always in the first argument – that justify incorporation of this characteristics as a separate element into the lexicon. The element ordering has one of these five values: *1* for connectives occurring only in the first argument, *2* for connectives in the second argument, *1 or 2* for connectives in the first or second argument, *1 and 2* for complex correlative connectives and *N/A* for secondary connectives forming a separate syntactic unit (e.g. *Důvod je jednoduchý.* [*The reason is simple.*]) and therefore occurring entirely between the arguments.
- element ***integration***: captures the position of the connective within the argument. According to their origin and other possible functions in text, Czech connectives have different positions in the argument. Only subordinating conjunctions and prototypical coordinating conjunctions occupy the very beginning of the clause or sentence; the position of other connectives varies. Some of them are placed typically at the clitic, i.e. second position (e.g. *však* [*however*]), some of them are typically either on the first or on the second position (e.g. *potom* [*then*] or *proto* [*therefore*]) and for the class of focusing particles (i.e. expressions like *také* [*also*] or *jenom* [*only*]), the position is given by the information structure. For secondary connectives represented by the whole clause, *integration* is again *N/A*. Other values of this element, as follows from examples just mentioned, are *first*, *second*, *first or second*, and *any*. For complex correlative connectives forming level-one entries, the value is given for the second part of the connective only.
- element ***realizations***: a list of non-modified and non-complex secondary connectives from PDiT 2.0 represented by the given dependency scheme (applies only to secondary connectives)
- element ***modifications***: a list of the connective modifications: e.g. for the lemma *potom* [*then*] expressing *precedence–succession*, there is a modification *teprve potom* [*only then*]. Secondary connectives can be modified as well – cf. *hlavní důvod proč* [*the main reason why*]. Modifications are further distinguished in the attribute *type* as *eval* (evaluative), *modal*, and *intense* (intensifying).
- element ***complex_forms***: a list of complex connectives: e.g. for the lemma *potom* [*then*] expressing *precedence–succession*, there are for example complex forms *a potom* [*and then*] and *nejdřív potom* [*first then*]. Secondary connectives can have

[23] This differs from the original design reported in Mírovský et al. (2016b) where this element signalled the linear order of the external argument. The new semantics of this element is more consistent with the semantics of elements *arg_semantics* and *integration*.

complex forms as well – cf. *a z tohoto důvodu* [*and for this reason*]. The criterion for a complex form to be placed in the level-two entry under a certain lemma is the ability of the basic connective (the given lemma) to express the same discourse type. It means that e.g. the complex connective *přesto však* [*yet however*] expressing the discourse type of *concession* is placed in respective level-two entries under both lemmas *přesto* [*yet*] and *však* [*however*], because both these single connectives individually also express the discourse type of *concession* in PDiT 2.0. Further, according to its placement either in both arguments or in one argument, each complex form is labeled in the attribute *type* as *correlative*, *continuous* or *discontinuous* (see above among the level-one entry characteristics).

- element ***examples***: a list of a few illustrative examples from PDiT 2.0 and their English translations. Both intra-sentential and inter-sentential examples are – if available in the corpus – given for the connective usages and marked as such in the attribute ***type*** (*intra* vs. *inter*).
- element ***is_rare***: signals a rare use of the connective with the given discourse type
- element ***register***: captures whether the connective is used in the *neutral*, *formal* or *informal* register
- attribute ***id***: a unique identifier of this level-two entry

For non-connective usages, the argument semantics, ordering, integration, modifications and complex forms are not applicable, whereas other characteristics are given similarly as for connective usages.

4.3.3. Corpus frequencies

Numbers of occurrences in PDiT 2.0 were added to all individual variants, complex forms, modifications and realizations, as well as to connective and non-connective usages (level-two entries) and the whole lemmas (level-one entries), in two attributes: ***pdt_count*** and ***pdt_intra***, capturing numbers of all vs. intra-sentential occurrences of the respective items.

Contrary to our former intention (stated in Mírovský et al., 2016b) to extract the lexicon from 9/10 of the source corpus only (leaving the last 1/10 of the data for test purposes), we decided in the end to use the whole PDiT 2.0 for the extraction, to have the whole data of the corpus covered and interconnected with the lexicon.[24] All numbers in the attributes *pdt_count* and *pdt_intra* therefore reflect frequencies from the whole PDiT 2.0.

[24] Similarly to e.g. PDT-Vallex, a lexicon of valency frames of verbs and (newly) some nouns in the Prague Dependency Treebank (see Urešová, 2011 and Kolářová, 2014), which also covers the whole treebank.

5. Practical Implementation

This section describes the implementation of the lexicon in the Prague Markup Language framework (PML, see Section 5.1 just below) and advantages this choice brings. We show details of the data format on several examples, to demonstrate a relative ease of using the PML formalism and possibly encourage others to use it in their practical research. We also describe steps in the process of extracting the lexicon from the Prague Discourse Treebank 2.0 and mention a few post-processing steps needed to improve the quality of the final data, and connective properties that needed to be inserted into the lexicon manually.

5.1. Prague Markup Language

The data format used in the Prague Discourse Treebank 2.0 is called the Prague Markup Language (PML, Hana and Štěpánek, 2012).[25] It is a data format used for many other treebanks developed in Prague or abroad, such as the Prague Dependency Treebank since version 2.0, the Prague Czech-English Dependency Treebank (Hajič et al., 2012), the Slovene Dependency Treebank (Džeroski et al., 2006), the Croatian Dependency Treebank (Berović et al., 2012), Ancient Greek and Latin Dependency Treebanks (Bamman and Crane, 2011), as well as all treebanks in the HamleDT project (Zeman et al., 2015), and many others.

The PML is an abstract XML-based format designed for annotation of richly linguistically annotated corpora, and especially treebanks. It is 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, including external lexicons.

The PML framework offers the following advantages:[26]

- 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.[27]
- 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 any data in the PML.

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

[26] The PML framework brings also low level tools for data validation (against a PML schema) and libraries to load and save data. And, of course, as the PML format is technically an XML, any general XML tool can be used for the data as well.

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

Using the PML framework presupposes representing the data in the PML format. Encoding a particular treebank in the PML requires:

- defining a PML-schema for each annotation layer of the data – this includes definition of tree node types, relations between the nodes, attributes for individual node types, values of the attributes,
- defining a stylesheet for the data – the stylesheet gives a full control over the way the data are displayed in the tree editor TrEd,
- and, optionally, defining macros – Perl scripts for manipulation with the data from within TrEd or btred; macros are often created to simplify the most common tasks done by the annotators.

The following listing 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 the definition of the format for level-one entries (the lemmas):

```
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="english"><cdata format="any"/></member>
07     <member name="type" type="c-type.type"/>
08     <member name="struct" type="c-struct.type"/>
09     <member name="variants" type="c-variants.type"/>
10     <member name="usages" type="c-usages-all.type" role="#CHILDNODES"/>
11   </structure>
12 </type>
```

Notice the declarations of roles (`role="#NODE"`, `role="#CHILDNODES"`, lines 2 and 10), 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.

Similar type definitions need to be provided for all other parts of the lexicon data structure, i.e. for the types referred to in the definition of the type `c-lemma.type` above and for all other data types needed in the lexicon. For example, the definition of the type `c-type.type` referred to from line 7 looks like this:

```
<type name="c-type.type">
  <choice>
    <value>primary</value>
    <value>secondary</value>
  </choice>
</type>
```

The following commented example shows the respective part of the resulting lexicon entry for the connective *potom* [*then, afterwards*]:

```
<lemma id="l-potom" pdt_count="95"> (a level-one entry)
  <text>potom</text>  (the lemma itself)
  <english>then; afterwards</english> (an approximate English translation;
    more precise translations are given at level-two entries)
  <type>primary</type>  (vs. secondary)
  <struct>single</struct>  (vs. complex)
  <variants>
    (no variants in the data for this lemma)
  </variants>
  <usages>
    <conn-usages pdt_count="80" pdt_intra="37">
      (list of connective usages, see Figure 2)
    </conn-usages>
    <non-conn-usages pdt_count="15">
      (list of non-connective usages)
    </non-conn-usages>
  </usages>
</lemma>
```

The commented example in Figure 2 shows a level-two entry in the PML for the lemma *potom* [*then, afterwards*], defining the lemma's connective usage with the semantic discourse type *precedence–succession*. The same part of the lexicon data is displayed in Figure 3 – it shows the lexicon loaded in the tree editor TrEd, allowing a user to inspect the record(s) or an annotator to make manual changes in the data. It displays the entry for the whole lemma, with an opened dialog window for editing the connective usage representing the discourse type *precedence–succession*, and a roll-down list of available options for the value of the element *arg_semantics*. The lemma (level-one entry), the list of connective usages, the list of non-connective usages, and the individual usages (level-two entries) are represented by tree nodes.

Using the PML for the lexicon brings, apart from the three advantages named earlier in this section, another possibility – the lexicon can be easily interlinked with the source data, i.e. the Prague Discourse Treebank 2.0, by adding identifiers of the lexicon entries (values of the attribute *id*, e.g. *c-potom-preced* from the example in Figure 2, line 1) 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:[28]

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

```
<usage id="c-potom-preced" pdt_count="63" pdt_intra="30">
  <sense>precedence-succession</sense>  (the represented semantic discourse type)
  <gloss>posléze</gloss>  (a synonym/explanation of the meaning in Czech)
  <english>afterwards</english>  (English translation)
  <pos>adverb</pos>  (part of speech)
  <arg_semantics>precedence-succession:succession</arg_semantics>
    (the argument associated with the connective represents
     the ``subsequent'' part of the relation)
  <ordering>2</ordering>  (the argument associated with the connective is placed
                           second in the surface order of the arguments)
  <integration>first or second</integration>  (a typical position in the argument)
  <register>neutral</register>  (vs. formal, informal)
  <modifications>  (a list of modifications)
    <modification type="intense" pdt_count="1" pdt_intra="1">
      <text>a teprve potom</text>  (an intensifying modification)
      <english>and only then</english>
    </modification>
  </modifications>
  <complex_forms>  (a list of complex forms)
    <complex_form type="discontinuous" pdt_count="14" pdt_intra="11">
      <text>a potom</text>
      <english>and then</english>
    </complex_form>
    (four more complex forms omitted to save space here)
  </complex_forms>
  <examples>  (a list of examples from PDiT 2.0)
    <example type="inter">  (an inter-sentential example)
      <text>Řekl sestře, že už nemůže dál, že si jde něco udělat, plakal
            a loučil se s ní. Potom odjel škodovkou.</text>
      <english>He told his sister that he could not go any further, that
            he was going to do something to himself, he cried and was saying
            goodbye to her. Then he drove away in his Škoda.</english>
    </example>
    <example type="intra">  (an intra-sentential example)
      <text>Psovod uvedl, že stopu pachatele ztratil a potom vyhledal jinou.</text>
      <english>The dog handler said that he had lost the perpetrator's trail
               and then found another.</english>
    </example>
  </examples>
  <pdt>  (information closely related to the source corpus)
    <discourse_type>preced</discourse_type>
    <pos_list>
      <pos>adverb</pos>
    </pos_list>
  </pdt>
</usage>
```

Figure 2. An abbreviated level-two entry for the lemma potom [then, afterwards] and the semantic discourse type precedence–succession.

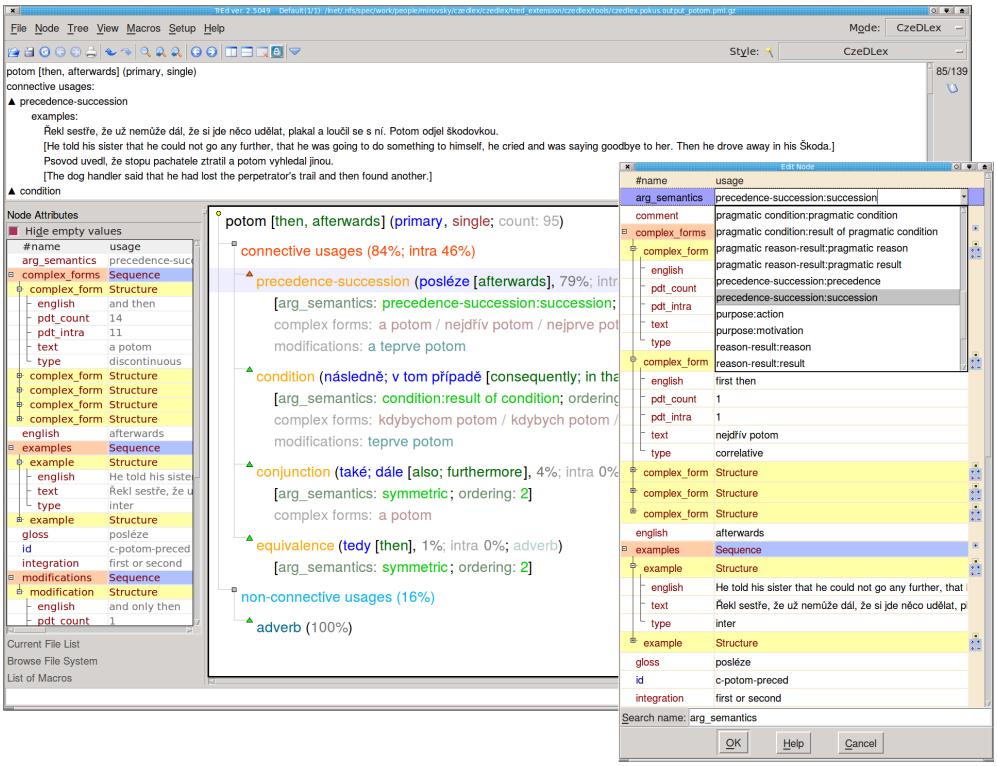

Figure 3. CzeDLex opened in the tree editor TrEd, with the lemma potom [then, afterwards] displayed. In the left panel, as well as in the pop-up window on the right side, information for the selected connective usage with the semantic discourse type precedence–succession is available. In the pop-up window, a pull-down menu for a selection of the argument semantics is being used.

- all occurrences of discourse relations in the treebank expressed by connectives that have the ability to express (in different contexts) more than X (e.g. 2) different discourse types (senses),
- all occurrences of discourse relations in the treebank expressed by connective words that are ambiguous in their connective vs. non-connective usages,
- all occurrences of discourse relations in the treebank expressed by complex or modified connectives.

5.2. Data Extraction

The process of extracting a raw base of the lexicon from the Prague Discourse Treebank 2.0 started with an extraction of a list of all connectives annotated in the treebank data, using a simple PML-TQ query. 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 already distinguished in the source corpus data and were treated separately. In over 20 thousand annotated discourse relations in the treebank, there were approx. 700 different items for the primary connectives and 350 for the secondary ones. Human annotators then manually divided the connectives into groups of connectives belonging to the same lemma, and in each group further distinguished complex forms, variants, modifications and (for the secondary connectives) realizations. For selected secondary connectives, also dependency schemes representing syntactically different realizations were created and the connectives were divided into subgroups according to the schemes.

This manually processed list served as an input for a btred[29] script that went through the whole data of the treebank, found all occurrences of the lemmas (and their variants, modifications etc.) and sorted them into the lexicon according to their type of usage (connective vs. non-connective) and the semantic discourse type of the relations (or the part of speech for non-connective usages). For each usage, a number of the shortest intra-sentential and inter-sentential examples[30] were collected (the annotators later chose the most suitable ones and added their English translations). Several other attributes could be set automatically as well – the part of speech, in most cases also the argument semantics and ordering (according to the orientation of the discourse arrow and position of the connective in an argument). Numbers of occurrences in PDiT 2.0 were added to all individual variants, complex forms and modifications, 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 were 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

[29] a command line version of the tree editor TrEd

[30] For some connectives, only one type of examples could be found. The distinction also does not apply to non-connective usages.

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 glosses, register, rareness, syntactic characteristics of secondary connectives) or the data were not big enough to cover all existing possibilities (dependency scheme, integration, sometimes ordering).

6. Conclusion

We have presented theoretical and implementation aspects of the design and development process of CzeDLex – a new electronic Lexicon of Czech Discourse Connectives. It is the first lexicon of Czech connectives and its uniqueness in the worldwide sense also lies in the fact that it includes not only primary but also secondary connectives. Special effort was dedicated to having both types of connectives represented in a relatively uniform way, as much as their different syntactic nature allows. We have also presented the data format used – the Prague Markup Language – and advantages this choice brings, and elaborated on the actual process of exploiting the source corpus, namely the Prague Discourse Treebank 2.0, to build the raw basis of the lexicon, with subsequent automatic and manual checks, corrections and additions.

Building the lexicon on the basis of an annotated corpus brings a certainty that the selection of the connectives and their coverage in the lexicon are to a certain degree representative but at the same time it sets limits on both these aspects, as the source treebank consists of newspaper texts only and, although it is large for a manually annotated treebank, its size is still limited.[31]

CzeDLex is built not only for theoretical purposes. Given its rich annotation of the properties of the connectives (including syntactic characteristics of the connectives, a general dependency scheme for the secondary connectives and distinction of variants, complex connectives and modified connectives), it may be useful also for NLP tasks that involve discourse parsing, for machine translation, information extraction and for text generation.

Our aim was also to make the lexicon readable for non-Czech speakers and to simplify its future interlinking with lexicons in other languages. We tried to achieve these goals by structuring the lexicon entries by semantic discourse types, by providing comprehensive morphological, syntactic and other characteristics both for the primary and secondary connectives, by using both human and computer readable format and by having all names of elements, attributes and their values (with the obvious exception of Czech word entries and Czech corpus examples) in English. In

[31] And much smaller than e.g. the SYN series of the Czech National Corpus, which contains automatically morphologically annotated texts in size of approx. 100 million words.

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[32] under the Creative Commons license. It will cover an essential part of the connectives used in the Prague Discourse Treebank 2.0.[33] The second version of CzeDLex, planned to be published next year, will cover all connectives annotated in the treebank.

Acknowledgements

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).

Bibliography

Al-Saif, Amal and Katja Markert. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *Proceedings of LREC 2010*, pages 2046–2053, Valletta, Malta, 2010.

Asher, Nicholas. *Reference to abstract objects in discourse*. Kluwer, Norwell, MA, 1993.

Ball, Wilson James. *Dictionary of link words in English discourse*. Macmillan, 1993.

Bamman, David and Gregory Crane. The ancient Greek and Latin dependency treebanks. In *Language technology for cultural heritage*, pages 79–98. Springer, 2011.

Bejček, Eduard, 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. Prague Dependency Treebank 3.0. Data/software, 2013.

Berović, Daša, Željko Agić, and Marko Tadić. Croatian dependency treebank: Recent development and initial experiments. In *Seventh International Conference on Language Resources and Evaluation (LREC 2012)*, 2012.

Breindl, Eva, Anna Volodina, and Ulrich Hermann Waßner. *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer*, volume 13. Walter de Gruyter GmbH & Co KG, 2015.

Buscha, Joachim. *Lexikon deutscher Konjunktionen*. Langenscheidt, Verlag Enzyklopädie, 1989.

[32] `http://lindat.cz`

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

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer, 2003.

Čermák, František. *Frazeologie a idiomatika: česká a obecná*. Karolinum, 2007.

Čermák, František. *Slovník české frazeologie a idiomatiky*. Leda, 2009.

Da Cunha, Iria, Juan-Manuel Torres-Moreno, and Gerardo Sierra. On the development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10. Association for Computational Linguistics, 2011.

Danlos, Laurence, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. Vers le FDTB: French Discourse Tree Bank. In *TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 471–478, 2012.

Džeroski, Sašo, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, and Andreja Žele. Towards a Slovene dependency treebank. In *Proc. of the Fifth Intern. Conf. on Language Resources and Evaluation (LREC)*, 2006.

Feltracco, Anna, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. LICO: A Lexicon of Italian Connectives. *CLiC it*, page 141, 2016.

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

Hajič, Jan, 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. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3153–3160, Istanbul, 2012. ELRA, European Language Resources Association.

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

Hausmann, Franz Josef. Lexikographie. *Handbuch der Lexikologie. Königstein: Athenäum*, pages 367–411, 1985.

Helbig, Gerhard. *Lexikon deutscher Partikeln*. Verlag Enzyklopädie, 1988.

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

Helbig, Gerhard and Agnes Helbig. *Lexikon deutscher Modalwörter*. Verlag Enzyklopädie, 1990.

Iruskieta, M., M. Aranzabe, A. Diaz de Ilarraza, I. Gonzalez, I. Lersundi, and O. Lopez de Lacalle. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop RST and Discourse Studies*, pages 40–49, Sociedad Brasileira de Computacao, Fortaleza, CE, Brasil, 2013.

Kolářová, Veronika. Valence vybraných typů deverbativních substantiv ve valenčním slovníku PDT-Vallex. Technical Report TR-2014-56, ÚFAL MFF UK, 2014.

Lin, Ziheng, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184, 2014.

Mann, William C. and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8:243–281, 1988a.

Mann, William C. and Sandra A. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281, 1988b.

Meyer, Thomas and Lucie Poláková. Machine translation with many manually labeled discourse connectives. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 43–50, Sofia, Bulgaria, 2013.

Meyer, Thomas, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Proceedings of the SIGDIAL 2011 Conference*, pages 194–203. Association for Computational Linguistics, 2011.

Mírovský, Jiří, Lucie Mladová, and Zdeněk Žabokrtský. Annotation Tool for Discourse in PDT. In Huang, Chu-Ren and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, volume 1, pages 9–12, Beijing, China, 2010. Chinese Information Processing Society of China, Tsinghua University Press.

Mírovský, Jiří, Pavlína Jínová, and Lucie Poláková. Discourse Relations in the Prague Dependency Treebank 3.0. In Tounsi, Lamia and Rafal Rak, editors, *The 25th International Conference on Computational Linguistics (Coling 2014), Proceedings of the Conference System Demonstrations*, pages 34–38, Dublin, Ireland, 2014. Dublin City University (DCU), Dublin City University (DCU).

Mírovský, Jiří, Lucie Poláková, and Jan Štěpánek. Searching in the Penn Discourse Treebank Using the PML-Tree Query. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1762–1769, Paris, France, 2016a. European Language Resources Association.

Mírovský, Jiří, Pavlína Synková, Magdaléna Rysová, and Lucie Poláková. 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, 2016b. Kyung Hee University, Kyung Hee University.

Oza, Umangi, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. The Hindi Discourse Relation Bank. In *Proceedings of the third Linguistic Annotation Workshop*, pages 158–161, 2009.

Pajas, Petr and Jan Štěpánek. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In Scott, Donia and Hans Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester, 2008. The Coling 2008 Organizing Committee.

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

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

Poláková, Lucie. *Discourse Relations in Czech*. PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, 2015.

Poláková, Lucie, Pavlína Jínová, and Jirí Mírovskỳ. Interplay of Coreference and Discourse Relations: Discourse Connectives with a Referential Component. In *LREC*, pages 146–153. Citeseer, 2012.

Poláková, Lucie, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, and Eva Hajičová. Manual for Annotation of Discourse Relations in Prague Dependency Treebank. Technical Report 47, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 2012a.

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

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

Poláková, Lucie, Pavlína Jínová, and Jiří Mírovský. Genres in the Prague Discourse Treebank. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1320–1326, Reykjavík, Iceland, 2014. European Language Resources Association.

Prasad, Rashmi and Harry Bunt. Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 80–92, 2015.

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, Philadelphia, 2007. URL `http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf`.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, 2008. European Language Resources Association.

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

Rysová, Magdaléna. *Diskurzní konektory v češtině (Od centra k periferii) [Discourse Connectives in Czech (From the Centre to the Perifery)]*. PhD thesis, Charles University, Prague, Czechia, 2015.

Rysová, Magdaléna and Kateřina Rysová. The Centre and Periphery of Discourse Connectives. In Aroonmanakun, Wirote, Prachya Boonkwan, and Thepchai Supnithi, editors, *Proceedings of Pacific Asia Conference on Language, Information and Computing*, pages 452–459, Bangkok, 2014. Department of Linguistics, Faculty of Arts, Chulalongkorn University, Department of Linguistics, Faculty of Arts, Chulalongkorn University.

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

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

Sanders, Ted JM, Wilbert PM Spooren, and Leo GM Noordman. Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35, 1992.

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

Schröder, Jochen. *Lexikon deutscher Präpositionen*. Verlag Enzyklopädie, 1986.

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

Stede, Manfred and Yulia Grishina. Anaphoricity in Connectives: A Case Study on German. *Coreference Resolution beyond OntoNotes*, page 41, 2016.

Stede, Manfred and Arne Neumann. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of LREC 2014*, pages 925–929, Reykjavik, Iceland, 2014.

Stede, Manfred and Carla Umbach. DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In *Proceedings of the 17th International Conference on Computational Linguistics (Coling 1998)*, pages 1238–1242. Association for Computational Linguistics, 1998.

Synková, Pavlína, Magdaléna Rysová, Lucie Poláková, and Jiří Mírovský. Extracting a Lexicon of Discourse Connectives in Czech from an Annotated Corpus. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 1–8, Cebu, Philippines, 2017, in print. University of the Philippines Cebu.

Urešová, Zdeňka. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011.

Urešová, Zdeňka, Eva Fučíková, and Jana Šindlerová. CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105:17–50, 2016.

Veselovská, Kateřina and Ondřej Bojar. Czech SubLex 1.0, 2013.

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

Zeyrek, Deniz and Murathan Kurfalı. TDB 1.1: Extensions on Turkish Discourse Bank. *LAW XI 2017*, page 76, 2017.

Zeyrek, Deniz, Işin Demirşahin, Ayişiği Sevdik-Çalli, Hale Ögel Balaban, İhsan Yalçinkaya, and Ümit Deniz Turan. The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In *Proceedings of the fourth Linguistic Annotation Workshop*, pages 282–289, 2010.

Zhou, Yuping and Nianwen Xue. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 69–77, 2012.

Zhou, Yuping and Nianwen Xue. The Chinese discourse treebank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397, 2015.

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

Address for correspondence:
Jiří Mírovský
mirovsky@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25
118 00 Praha 1
Czech Republic