Designing CzeDLex – A Lexicon of Czech Discourse Connectives

Jiří Mírovský, Pavlína Synková, Magdaléna Rysová, Lucie Poláková

Charles University in Prague, Czech Republic
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{mirovsky|synkova|magdalena.rysova|polakova}@ufal.mff.cuni.cz

Abstract

We present a design for a new electronic lexicon of Czech discourse connectives. The data format and the annotation scheme are based on a study of similar existing resources, and we discuss arguments for choosing the data structure and selecting features of the lexicon entries. A special attention is paid to a consistent encoding of both primary and secondary connectives. The data itself comes from exploiting the Prague Dependency Treebank, a large treebank manually annotated with discourse relations.

1 Introduction

Electronic lexicons of discourse markers[1] are not only a useful tool in theoretical research of text coherence/cohesion, they may also help in practical tasks such as discourse parsing, disambiguation of non-connective vs. connective usages of discourse markers, determining semantic type of discourse relations the markers convey, and also in selecting the correct counterpart of a discourse marker in translation from one language to another. Generally, systematic information on discourse markers contributes to processing discourse relations and in that way it helps NLP applications such as machine translation, text generation, information extraction, and others (cf. e.g. Meyer et al. (2011), Stede (2014) or Lin et al. (2014)).

Our goal has been to design and create an electronic lexicon of Czech 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 in other languages.

Once an annotation scheme of a lexicon is established, there are several options how to actually build the lexicon, i.e. how to fill it with data, from consulting existing printed lexicons, to using translation from lexicons in other languages or even just parallel texts, to exploiting existing (preferably discourse-annotated) corpora in the given language. We have chosen the last option, as a large discourse-annotated treebank – the Prague Dependency Treebank (see Section 1.2) – is available for Czech, and we are currently in the process of entering the data into the lexicon.

The subsequent text is organized as follows: First, in the rest of Introduction, we give an overview of main existing electronic lexicons of discourse markers that served as inspiration for our own work, and describe shortly the Prague Dependency Treebank with focus on its dis-

[1] We use "discourse markers" as a broader term for expressions generally structuring discourse, and "discourse connectives" as a narrower term for expressions signalling semantico-pragmatic relations between two abstract objects – see Section 2.1.

course annotation. Section 2 starts with delimiting the class of expressions we are interested in, i.e. the definition of connectives, their division into primary and secondary ones, and how we understand the terms compound connectives and modified connectives. We discuss issues related to nesting all these types of connectives in the lexicon (including their non-connective usages), as well as issues related to placement of connectives in their arguments. In Section 3, we discuss the selection of data for the lexicon and present the data format and annotation scheme for CzeDLex on two examples, and then we conclude in Section 4.

1.1 Existing Lexicons

Nowadays there are many corpora annotated with discourse relations but electronic lexicons of discourse connectives are much less common. We mention the most important ones.

DiMLex is a lexicon of German discourse markers; it was first introduced in Stede and Umbach (1998) and Stede (2002)(by then it was focused on syntactic properties of the connectives) and recently updated (Scheffler and Stede, 2016) with the annotation of discourse types – senses – from the Penn Discourse Treebank 3.0 (PDTB 3.0) sense hierarchy.[2] It is a computer-oriented resource, encoded in XML, with the main practical purpose to help NLP applications such as text generation and text understanding. It currently covers 275 German connectives.

LexConn is a lexicon of French discourse connectives (Roze et al., 2012), consisting of 328 connectives with their morphological categories, syntactic properties and discourse relations they convey according to the SDRT framework (Asher and Lascarides, 2003). Similarly to DiMLex, it is also a computer-oriented resource, encoded in XML, with the main practical purpose to help in NLP tasks that involve discourse parsing.

DPDE (Diccionario de partículas discursivas del español) is a dictionary of Spanish discourse markers (Briz et al., 2003). It consists of 229 entries and for each of them, it provides a definition, translation, ambiguous meanings, prosody, position, syntax, partial synonyms, idioms, register, and non-DM uses. Given its format (only HTML online) and annotation scheme (properties of markers are defined in plain language), its purpose is mostly for theoretical research.

1.2 Prague Dependency Treebank

The Prague Dependency Treebank (PDT) represents a richly annotated corpus with a multilayer annotation of approx. 50 thousand sentences of Czech newspaper texts. It contains morphological information and two layers of syntactic annotation, both of them in a form of dependency trees.

Annotation of discourse relations in the PDT was carried out on top of the deep-syntax layer (tectogrammatics) and covers explicit discourse relations, i.e. discourse relations anchored by a surface present connective. For sense annotation, a modified version of the PDTB 2.0 taxonomy was used, see Zikánová et al. (2015). The annotation proceeded in two phases. The first phase involved primary connectives (expressions like *však* [*however*] or *proto* [*therefore*]), arguments of these relations were limited to structures governed by a finite verb (mainly clauses or sentences). This annotation was published in the PDT 3.0 (Bejček et al., 2013). In the second phase, secondary connectives (expressions like *z toho důvodu* [*for that reason*], *to znamená* [*it means*]) were covered; the annotation of secondary connectives involved also relations with nominal phrases as arguments. Its publication is in process.

2 Theoretical Aspects

A crucial issue for building a lexicon of discourse connectives is a delimitation of this category. Since CzeDLex is based on the annotation of discourse relations in the PDT, it adopts also the PDT approach to discourse connectives.

2.1 Theoretical Delimitation

In the PDT, a discourse connective is defined as a predicate of a binary relation opening two positions for two text spans as its arguments and

[2] The PDTB 3.0 sense hierarchy is to be published later this year, for the PDTB 2.0 sense hierarchy see e.g. Prasad et al. (2008).

signalling a semantic or pragmatic relation between them.[3]

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, over simple clauses to participial and infinitive constructions and nominal phrases. In the PDT, annotation of discourse arguments was syntactically restricted to verbal arguments (i.e. whose basis is a finite verb).[4] The same restriction has been adopted also for CzeDLex.

Primary and secondary connectives
Discourse connectives in the PDT are divided into primary and secondary, according to Rysová and Rysová (2014). They differ especially in the degree of their grammaticalization. Primary connectives are rather short and grammaticalized expressions belonging to certain parts of speech (mostly conjunctions, particles and some types of adverbs), such as (in English[5]) *while*, *because*, *however*, *therefore*. On the other hand, secondary connectives are especially multiword phrases like *for this reason*, *to follow*, *due to this* etc. that are not yet fully grammaticalized (see also Rysová and Rysová (2015)).[6]

Since the PDT contains detailed annotation of both primary and secondary connectives, we include both of these types also into CzeDLex. However, primary and secondary connectives differ in many important aspects that need to be reflected in the lexicon design: lemmatization, syntactic characteristics, part-of-speech appurtenance, placement of the external argument and argument integration (i.e. placement of a connective in the argument).

Generally, the difficulty of secondary connectives is that many of them 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 equal). See Sections 2.2, 2.4 and 2.5 below.

Compound and modified connectives
Discourse connectives may be further divided into the following categories: compound vs. single and modified vs. non-modified. Compound connectives consist of two or more connective words all participating on expressing the given discourse relation type. Compound connectives occur in a single argument (*a proto* [*and therefore*]) or they may form correlative pairs (*buď_nebo* [*either_or*]). A compound connective may express the same or different semantic type than its individual parts.

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

To sum up, all members of compound connectives participate on expressing the particular discourse relation (e.g. both parts of *a proto* [*and therefore*] express together a relation of reason–result, and both parts of *buď_nebo* [*either_or*] express a relation of disjunctive alternative), while in modified connectives, the modification (e.g. *mainly* in *mainly because*) does not participate on expressing a discourse relation type (in our example reason–result) but it only modifies it (it expresses the intensity of the relation). For more details see M. Rysová (2015).

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. The non-connective uses of these homonymous expressions can be categorized into several groups with specific properties:

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

[4] with the exception of secondary connectives

[5] For simplicity, in the subsequent text we often present – when it is sufficient – only English equivalents of Czech expressions.

[6] The annotation and description of primary connectives in the PDT is given in detail in Poláková (2015) and of secondary connectives in M. Rysová (2015).

- From the perspective of a discourse analysis defining a discourse argument as an abstract object (Asher, 1993), expressions connecting mere entities (*mum* ***and*** *dad*) are not considered discourse connectives.
- Expressions in the role of expressive particles, reaching almost the role of interjections, are not treated as discourse connectives. They may function in discourse structuring, possibly within the wider category of discourse markers but they do not connect two abstract objects in our sense (***Tak*** *co s tím, nová rado?* [***So*** *what (do you do) about it, new council?*]).
- Expressions (homonyms of primary connectives) in the role of sentence constituents, mostly moreover in the rhematic part of the sentence, are not considered connectives[7] (*Vana plechová se zahřeje rychle a rychle zchladne, vana litinová se chová* ***naopak****.* [*A metallic bathtub gets heated quickly and quickly cools, a cast-iron bathtub behaves* ***otherwise****.*]).
- Expressions functioning as answer particles are not considered connectives (*Půjdeš tam?* ***Ovšem****.* [*Will you go there?* ***Of course****.*])

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 Section 2.2 just below), along with their syntactic characteristics.[8]

2.2 Nesting

On the first level of the CzeDLex structure, entries are nested according to a lemma of the connective.[9]

Lemma for secondary connectives
Selecting a representative lemma for primary connectives is a straightforward decision but for secondary connectives, a suitable similar approach needs to be found. For example, there are many secondary connectives containing the word "reason" (*for this reason*, *that is the reason why*, *the reason is* etc.), and we consider the word "reason" their common "core" word. In our approach, we extract these "core" words of secondary connectives, 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.), and treat these "core" words (their lemmas) as umbrella lemmas for all individual variants.

Level-two nesting
There are two main options for nesting level-two entries in the lexicon – according to a lemma with a PoS tag (this principle has been adopted e.g. in DiMLex) or according to a lemma combined with a discourse semantic type (similarly to LexConn). For CzeDLex, we have chosen the latter option for these reasons: (i) part-of-speech annotation of discourse connectives in the PDT is outdated, (ii) part-of-speech appurtenance for connectives and expressions homonymous with them is often questionable and (iii) in machine-translation systems, links between lexicon entries in the involved languages need to be tied to discourse semantic types (the same preference comes also from text generation tasks).

If we followed this rule strictly, the depth of the lexicon scheme for secondary connectives would increase to three levels, as secondary connectives usually form several different syntactic structures (which need to be captured in separate entries), while still conveying the same semantic discourse type. To keep the scheme of the lexicon simpler and more unified for primary and secondary connectives, we cluster the level-two entries for secondary connectives not only by the semantic discourse type but also by the syntactic structure of similar surface realizations

[7] Secondary connectives (or their parts) are always sentence constituents (in contrast to primary ones). However, their "core" words may also have a non-connective function – cf. *the suggestion was rejected for procedural* ***reasons****.*

[8] A detailed analysis of "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).

[9] We use the morphological lemma rather than the tectogrammatical lemma, as many connectives are not represented as nodes on the tectogrammatical layer of the PDT (thus they do not have a tectogrammatical lemma).

of the connective.

To describe all possible realizations (in the PDT) of a secondary connective that conform to the same syntactic structure (and thus fall into the same lexicon entry), we establish a general pattern for such a structure, expressed by a linear text notation of the dependency representation of the structure on the surface-syntax layer of the PDT – see e.g. the scheme *(anaph. Sb) Pred ([(Atr)] důvod.1,7) AuxC*[10] for realizations such as *to je důvod, proč*; *to byl hlavní důvod, proč*; *to je důvodem, proč* [all meaning *that is/was the (main) reason why*]. See another example in the XML element *schema_dep* in Section 3.2.2.

PoS for secondary connectives
Another issue concerns the part-of-speech appurtenance of discourse connectives. Whereas we may relatively easily define the part of speech for primary connectives, the situation with secondary connectives is less simple, as they form whole syntactic structures (like *under these conditions*). At level one of the lexicon, we only define the part-of-speech category of the "core" word (i.e. of the lemma), and for each individual variant of the secondary connective (represented at level two), we state the global syntactic characteristics of the whole expression (e.g. *under these conditions* – prepositional phrase), see the XML element *syntactic_characteristics* in Section 3.2.2.

Compound and modified connectives
Single connectives (such as *a* [*and*], *ale* [*but*], *protože* [*because*]), in combination with their individual semantic types, are in the lexicon always treated as separate entries.

Within compound connectives, only those expressing a semantic type different from those expressed by the members of the compound connective themselves have a separate entry (e.g. *i když* [lit. *also if*, meaning *even though*][11]). Other compound connectives (like *a proto* [*and therefore*], *i proto* [lit. *also therefore*]) are listed under a semantically "stronger" connective (e.g. *proto* [*therefore*]). Including compound connectives into the lexicon is important for NLP applications, as processing them separately by the individual parts might lead to incorrect results.

Modified connectives are not treated as separate entries in CzeDLex, as they do not change the semantic discourse type. Instead, the modifications (that occur in the PDT) are listed under the relevant non-modified connective.

2.3 Semantics of Arguments

From the semantic point of view, there is a difference between symmetric and asymmetric discourse relations. Whereas for symmetric relations, the general semantic characteristics is shared by both arguments, asymmetric discourse relations hold between arguments that reveal different semantic characteristics. For example, if the arguments are in the asymmetric relation of reason–result, one of them expresses a reason, the other one a result.[12]

Typically, a connective is characterized by its placement in one specific part of the relation it signals. For example, coordinating conjunction *tedy* [*thus*] signals a result, while *totiž* [*because*] signals a reason. In CzeDLex, this characteristics of connectives in asymmetric relations is captured in the XML element *arg_semantics* (see Section 3.2).

2.4 Position of the External Argument

A connective and its position not only help determine the semantics of the arguments (and the whole relation), but also positions of the arguments. This characteristics is given by part-of-speech appurtenance for almost all primary connectives in Czech. Coordinating conjunctions, adverbs and particles are placed in the

[10] Where *anaph. Sb* means an anaphoric subject, *Pred* is a predicate, *Atr* is an attribute, *důvod.1,7* means the word *důvod* [*reason*] in nominative or instrumental, *AuxC* is a subordinating conjunction; elements in square brackets [ ] are optional, parentheses ( ) mark syntactic dependencies.

[11] The single primary connective *i* [*also*] signals mostly a conjunction, *když* [*if*] signals mostly a condition and together they express a relation of concession.

[12] This (a)symmetry has to be addressed one way or another in any approach to discourse relations (see e.g. Prasad and Bunt, 2015; Sanders et al., 1992; Prasad et al., 2007).

linearly second argument (so the external argument has to be searched for in the previous text), while subordinating conjunctions are not specific in this respect – the external argument can be placed before or after the clause with a subordinating conjunction. There are however exceptions to this rule, for example the connective particle *nejenže* [lit. *not only that*] always signals the linearly first argument of the gradation relation.

It is therefore important to capture information about placement of the external argument of the relation in the lexicon. In CzeDLex, the XML element is called *ordering* (the label has been adopted from DiMLex) and carries a value expressing whether the external argument is in the previous context, the later context, or that both options are possible.[13]

2.5 Placement of a Connective in an Argument

According to their origin and functions, Czech connectives are placed at different positions in the argument. Only subordinating conjunctions and several prototypical coordinating conjunctions are placed at the beginning of a clause or sentence, but mostly, Czech connectives are placed elsewhere. Some of them obligatorily at the clitic, i.e. second position of the sentence (e.g. *však* [*however*]), but the position can also vary between first and second (e.g. *proto* [*therefore*] or *ale* [*but*]). A specific case is represented by so called focus particles, which signal the focus of the sentence and their placement thus varies according to information structure of the sentence.

The placement of the connective in an argument is captured in the XML element *integration* (the name taken again from DiMLex), with values for "first", "second", "first or second" or "any" position, and also "N/A" (non-applicable). The last value is used for secondary connectives represented by a whole clause.

[13] There is a special (fourth) value for those types of secondary connectives that occur entirely between arguments as a separate syntactic unit (like *Důvod je jednoduchý.* [*The reason is simple.*]).

3 CzeDLex

3.1 Data Selection

Entries for the Lexicon of Czech Discourse Connectives (CzeDLex) are being selected on the basis of the Prague Dependency Treebank, a corpus annotated with discourse relations (see Section 1.2). For the first version of CzeDLex, approx. 100 most common connectives will be processed. As the lexicon is intended to be used in NLP tasks, throughout the whole process – from designing the lexicon to selecting the connectives and their semantic types – we only use 9/10 of the PDT, leaving the predefined etest data unseen for allowing correct testing of applications that will use the lexicon in the future.

3.2 CzeDLex Annotation Scheme

The annotation scheme for the lexicon of Czech connectives is presented in this section on two commented examples: one primary connective and one secondary connective. For space restrictions, less important parts have been left out. We have chosen XML as the data format, following the examples of DiMLex and LexConn; it also simplifies integration into the PDT annotation framework (Pajas and Štěpánek, 2008).

3.2.1 A Primary Connective

The following is a shortened schema for a lexicon entry of a primary connective, demonstrated on the connective *tedy* [*so, therefore*].

```
<lemma id="l-tedy"> (a level-one entry)
  <text>tedy</text> (the lemma itself)
  <type>primary</type> (vs. secondary)
  <struct>single</struct>
      (vs. continuous, discontinuous, correlative)
  <variants>
    <variant register="informal">teda</variant>
  </variants>
  <connective_usages>
    (list of connective usages, see below)
  </connective_usages>
  <non-connective_usages>
    (list of non-connective usages)
  </non-connective_usages>
</lemma>
```

One of the connective usages is described in the following example. The discourse type repre-

sented by this level-two entry is reason–result.

```
<connective_usage id="c-tedy-reason">
  <discourse_type>reason-result</discourse_type>
  <gloss>proto</gloss>
  <english>so, therefore</english> (English transl.)
  <pos>conjunction</pos> (part of speech)
  <subpos>coord</subpos> (a detailed POS)
  <arg_semantics>result</arg_semantics>
    (the argument associated with the connective
     represents the ''result'' part of the relation)
  <ordering>2</ordering>
    (the (same) argument is always second in the text)
  <integration>first or second</integration>
    (position in the argument)
  <modifications> (list of modifications)
    (N/A for ''tedy'')
  </modifications>
  <compounds> (list of compounds)
     <compound struct="discontinuous">
      <orig>a tedy</orig>
      <english>and therefore</english>
     </compound>
  </compounds>
  <examples>
    (list of a few examples from the PDT, see below)
  </examples>
  <usage>standard</usage> (vs. rare)
  <register>neutral</register> (vs. e.g. formal)
  <pdt>
    (PDT-related info, e.g. POS and sub-POS according
     to the PDT, statistics etc.)
  </pdt>
</connective_usage>
```

The following is a slightly shortened PDT example of a connective usage of lemma *tedy* with discourse type reason–result:

```
<example>
  <orig>Přitom právě u dlouhodobějších investic,
    jako je stavba bytových domů, právní nejistota
    výrazně zvyšuje úroky z úvěrů, čímž snižuje
    nabídku. Legislativní činnost je tedy
    nejlevnější cestou, jak nabídku stimulovat, ...
  </orig>
  <english>
    But especially in long-term investments such as
    the construction of residential houses, legal
    uncertainty significantly increases the interest
    on loans, thereby reducing the supply.
    A legislative action is therefore the
    cheapest way to stimulate the supply.
  </english>
</example>
```

3.2.2 A Secondary Connective

The following is a shortened schema of a lexicon entry for a secondary connective, demonstrated on a connective with the core word *důvod* [*reason*]. The level-one entry is almost identical to a level-one entry of a primary connective, with the exception of the element *struct*, which – for secondary connectives – has been moved to level-two entries.

```
<lemma id="l-důvod">
  <text>důvod</text>
    (a lemma of the core word of the secondary conn.)
  <type>secondary</type> (vs. primary)
  <pos>noun</pos> (PoS of the core word)
  <connective_usages>
    (list of connective usages, see below)
  </connective_usages>
  <non-connective_usages>
    (list of non-connective usages)
  </non-connective_usages>
</lemma>
```

One of the connective usages is described in the following example. The discourse type represented by this level-two entry is reason–result. As for secondary connectives, there may be several level-two entries for the same discourse type, the identifiers (attribute id) carry a suffix number (-1, -2, etc.). Again, the level-two entry is almost identical to a level-two entry of a primary connective, with these exceptions: the *struct* element has been moved here from the level-one entry, part-of-speech elements have been replaced by *syntactic_characteristics* and *schema_dep* and complemented by the *realizations* element, which gives the most frequent examples of actual secondary connectives described by the dependency schema.

```
<connective_usage id="c-důvod-reason-1">
  <discourse_type>reason-result</discourse_type>
  <gloss>proto</gloss>
  <english>therefore</english>
  <syntactic_characteristics>
    prepositional phrase
  </syntactic_characteristics>
  <schema_dep>
    z ((anaph. Atr) důvod.2)
  </schema_dep>
```

```
<realizations>
  <realization>
    <orig>z tohoto důvodu</orig>
    <english>for this reason</english>
  </realization>
  <realization>
    <orig>z uvedených důvodů</orig>
    <english>for the stated reasons</english>
  </realization>
</realizations>
<struct>single</struct>
<arg_semantics>result</arg_semantics>
<ordering>2</ordering>
<integration>any</integration>
<modifications>
  <modification_type="eval">
    <orig>z tohoto prostého důvodu</orig>
    <english>for this simple reason</english>
  </modification>
  <modification_type="modal">
    <orig>z tohoto možného důvodu</orig>
    <english>for this possible reason</english>
  </modification>
</modifications>
<compounds>
  <compound struct="discontinuous">
    <orig>a z tohoto důvodu</orig>
    <english>and for this reason</english>
  </compound>
</compounds>
<examples>
  (list of a few examples from the PDT, see below)
</examples>
<usage>standard</usage> (vs. rare)
<register>neutral</register> (vs. e.g. formal)
<pdt>(PDT-related info, statistics)</pdt>
</connective_usage>
```

And a slightly simplified PDT example:

```
<example>
  <orig>S ohledem na toto ustanovení by se hrubé
    chování muselo týkat vaší osoby a nestačí pouze
    nevhodné zacházení s předmětem darovací smlouvy.
    Z tohoto důvodu by byla vaše žaloba na vrácení
    daru u soudu zamítnuta.
  </orig>
  <english>With regard to this regulation, the rough
    behaviour would have to involve your person and
    not simply improper handling of the subject of
    the donation agreement. For this reason, your
    legal action on the return of the donation with
    the court would be rejected.
  </english>
</example>
```

4 Conclusion

We have presented the design of CzeDLex – a Lexicon of Czech Discourse Connectives. It is the first lexicon of Czech connectives and its uniqueness also lies in the fact that it includes secondary connectives (existing lexicons of connectives for other languages do not cover expressions like *for this reason*, *to conclude* etc.).

We are currently in the process of filling the lexicon with data. The first version of CzeDLex will contain approx. 100 most frequent Czech discourse connectives according to the annotation of discourse relations in the PDT. Building the lexicon on the ground of a discourse-annotated corpus brings a certainty that the selection of the connectives and their coverage in the lexicon is to a certain degree representative but at the same time it sets limits on both these aspects, as the treebank consists of newspaper texts only and, although it is large for a manually annotated treebank, its size is still limited.

CzeDLex is built not only for theoretical purposes. Given its rich annotation of the properties of the connectives (including the general scheme for secondary connectives and inclusion of compound connectives), it may be useful also for NLP tasks that involve discourse parsing, for machine translation, and for text generation.

Our aim was also to make the lexicon readable for non-Czech speakers, and simplify its possible interlinking with lexicons in other languages, which we try to achieve by using both human and computer readable format and by providing English equivalents for all Czech entries, and also by providing comprehensive morphological, syntactic and other characteristics both for the primary and secondary connectives.

Acknowledgments

The authors gratefully acknowledge support from the Ministry of Education, Youth and Sports of the Czech Republic (projects COST-cz LD15052 and Kontakt LH14011). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry (project LM2015071).

References

Nicolas Asher. 1993. *Reference to Abstract Objects in Discourse.* Kluwer Academic Publishers.

Nicolas Asher and Alex Lascarides. 2003. *Logics of Conversation.* Cambridge University Press.

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, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czechia, http://ufal.mff.cuni.cz/pdt3.0/.

Antonio Briz, Salvador Pons Bordería, and José Portolés. 2003. *Diccionario de partículas discursivas del español.* Data/software, www.dpde.es. Online since 2003.

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

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

Petr Pajas, Jan Štěpánek. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In: *Proceedings of the 22nd International Conference on Computational Linguistics*. Coling 2008 Organizing Committee, Manchester, UK.

Lucie Poláková 2015. *Discourse Relations in Czech.* Ph.D. Thesis. Charles University in Prague, Czechia.

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

Rashmi Prasad and Nikhil Dinesh and Alan Lee and Eleni Miltsakaki and Livio Robaldo and Aravind Joshi and Bonnie Webber. The Penn Discourse Treebank 2.0. *Proceedings of LREC 2008*, pp. 2961–2968, Marrakech, Morocco.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association, Marrakech, 2961–2968.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L. Webber. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. Philadelphia: University of Pennsylvania.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: A French Lexicon of Discourse Connectives. *Discours [En ligne], 10|2012*, http://discours.revues.org/8645.

Magdaléna Rysová and Kateřina Rysová. 2014. The Centre and Periphery of Discourse Connectives. In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. Bangkok, Thailand, 452–459.

Magdaléna Rysová and Kateřina Rysová. 2015. Secondary Connectives in the Prague Dependency Treebank. In: *Proceedings of the 3rd International Conference on Dependency Linguistics (Depling 2015)*. Uppsala University, Sweden, 291–299.

Magdaléna Rysová. 2015. *Diskurzní konektory v češtině (Od centra k periferii)* [*Discourse Connectives in Czech (From Centre to Periphery)*]. Ph.D. Thesis. Charles University in Prague, Czechia.

Ted Sanders, Wilbert Spooren, and Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes. 15, 1*, Taylor & Francis.

Tatjana Scheffler and Manfred Stede. 2016. Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association, Paris, France.

Manfred Stede and Carla Umbach. 1998. DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding In: *Proceedings of the 17th International Conference on Computational linguistics.* pp. 151–184. Association for Computational Linguistics.

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

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

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*, Prague: ÚFAL, Charles University in Prague.