Describing CzeDLex – a Lexicon of Czech Discourse Connectives

Magdaléna Rysová, Lucie Poláková, Jiří Mírovský, Pavlína Synková

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

Abstract. In the present contribution, we introduce a pilot version of CzeDLex, a Lexicon of Czech Discourse Connectives. Currently, CzeDLex contains 205 lemmas of connectives coming from the annotation of the Prague Discourse Treebank 2.0 (PDiT). CzeDLex reflects division of connectives into primary (e.g. *když* [if]) and secondary (e.g. *za této podmínky* [under this condition]). Altogether, 134 lemmas in CzeDLex are primary connectives and 71 are lexical cores of secondary connectives (i.e. words like *podmínka* [condition]). All 205 lemmas are manually annotated with basic linguistic information; the full annotation is now in progress. At this stage, 19 lemmas have been fully manually processed, which covers more than two thirds of all discourse relations in the PDiT. In the present paper, we describe the process of building CzeDLex, we give a list of connective properties annotated in lexicon entries of both primary and secondary connectives and we present the way of their nesting. The technical solution of CzeDLex is based on the (XML-based) Prague Markup Language that allows for an efficient incorporation of the lexicon into the family of Prague treebanks and also for interconnecting CzeDLex with existing lexicons in other languages.

Keywords: CzeDLex, Discourse Connectives, Lexicon.

1 Introduction

In the present contribution, we introduce a pilot version of CzeDLex (a Lexicon of Czech Discourse Connectives, developed within the COST-cz project TextLink-cz). The lexicon is a result of a long-term investigation of Czech discourse relations in both theoretical and practical aspects (see e.g. the monograph by Zikánová et al., 2015; summarizing research of coherence with focus on discourse relations in Czech) and logical follow-up of Prague annotation projects like Prague Discourse Treebank 1.0 (PDiT, see Poláková et al., 2012) and 2.0 (Rysová et al., 2016) – a large corpus annotated with discourse relations and discourse connectives. CzeDLex is thus based on an extensive linguistic research of discourse in Czech.

CzeDLex contains connectives partially automatically extracted from the PDiT 2.0. The lexicon entries are being manually checked and supplemented by additional lin-

guistic information, starting with the most frequent connectives. The current development version of the lexicon is available online (http://ufal.mff.cuni.cz/czedlex/) and was published as a pilot version (version 0.5, Mírovský et al., 2017) in the Lindat/Clarin repository.

The data format and the data structure of the lexicon are based on a study of similar existing resources, especially on DiMLex – a lexicon of German discourse markers first introduced by Stede and Umbach (1998) and Stede (2002) and recently updated by Scheffler and Stede (2016). The main principle adopted for nesting entries in CzeDLex is a semantic type of discourse relations expressed by the given connective word, which enables us to deal with a broad formal variability of connectives. The technical solution of CzeDLex is based on the (XML-based) Prague Markup Language that allows for an efficient incorporation of the lexicon into the family of Prague treebanks – it can be directly opened and edited in the tree editor TrEd (see Pajas and Štěpánek, 2008), processed from the command line in btred, interlinked with its source corpus and queried in the PML-Tree Query engine (details on PML-TQ are given in Štěpánek and Pajas, 2010) – and also for interconnecting CzeDLex with existing lexicons in other languages.

In this presentation, we first discuss theoretical linguistic aspects underlying the division and the description of Czech connectives adopted in CzeDLex, we present a list of connective properties annotated in the lexicon and finally, we provide an example of a lexicon entry (the connective *proto* [therefore]).

2 Theoretical Linguistic Aspects behind CzeDLex – Division of Connectives

CzeDLex reflects a division of discourse connectives into primary and secondary (the terms and definitions introduced by Rysová and Rysová, 2014) which differ especially in the degree of 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 English *but*, *or*, *when*, *thus*. On the other hand, secondary connectives are especially multiword phrases like *under these conditions*, *this means*, *because of this* etc. that are not yet fully grammaticalized. At the same time, secondary connectives contain the so-called core words, cf. e.g. the word *condition* in structures like *under this condition* or *on condition that* etc. (see also Rysová and Rysová, 2015).[1] Since the PDiT 2.0 contains a detailed annotation of both primary and secondary connectives, both of these types are included also into CzeDLex.

Discourse connectives in CzeDLex are further divided into the following categories: complex vs. single and modified vs. non-modified (Rysová, 2015). Complex connectives consist of two or more connective words all participating in expressing the given discourse relation type. Complex connectives occur in a single argument (*a*

[1] The annotation and description of primary connectives in the PDiT is given in Poláková (2015) and of secondary connectives in Rysová (2015).

proto [and therefore]) or they may form correlative pairs (*buď_nebo* [either_or]). 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]).

3 List of Connective Properties in CzeDLex

3.1 Level-One and Level-Two Entries

The entries in CzeDLex are structured according to a two-level nesting principle. On the first level, entries are nested according to the lemma of a connective and contain the following linguistic information:

• type of the connective (primary vs. secondary),
• structure of the connective (single vs. complex),
• variants of the connective (e.g. stylistic or orthographic),
• connective usages – a list of level-two entries representing semantico-pragmatic relations the connective expresses and their properties,
• non-connective usages – another list of level two entries, representing contexts where the lemma does not function as a discourse connective (e.g. *young and beautiful*).

Level two for primary connectives reflects the discourse-semantic types (usages) and contains the following pieces of information:

• semantic type of the discourse relation (condition, opposition etc.),
• gloss (an explanatory Czech synonym),
• English translation,
• part of speech of the connective,
• argument semantics (for asymmetric relations like reason–result, e.g. *protože* [because] expresses reason while *proto* [therefore] expresses result),
• ordering, i.e. position of the argument syntactically associated with the connective in relation to the other (external) argument,
• integration, i.e. placement of a connective in an argument,[2]
• list of connective modifications,
• list of complex connectives containing the given connective,
• examples from the PDiT (i.e. a context for the given discourse relation) and their English translations,
• is rare (set to '1' for rare usages),
• register (formal, neutral, informal).

[2] The names of the elements ordering and integration are taken from DiMLex (Scheffler and Stede, 2016).

An entry for a secondary connective contains several modifications. On level one of the lexicon structure, entries are nested according to the lemma of the core word for a secondary connective (see above). A level-two entry then contains the following additional properties (details on them are given in Rysová, 2015):

• syntactic characteristics of the structure (e.g. *za této podmínky* [under this condition] is a prepositional phrase),
• dependency scheme (general pattern) for each structure (e.g. *za této podmínky* [under this condition] = "*za* ((anaph. Atr) *podmínka*.2)", i.e. a preposition *za* [under] plus an anaphoric attribute and the word *podmínka* [condition] in genitive),
• realizations of the dependency scheme (e.g. *za této podmínky* [under this condition], *za dané podmínky* [under the given condition] etc.).

Details on building and designing of CzeDLex are given in detail in Mírovský et al. (2016), Synková et al. (2017) and Mírovský et al. (2017).

3.2 Frequencies from the PDiT 2.0

The lexicon entries are also enriched by frequencies of the individual connectives in the PDiT 2.0. Numbers of occurrences in the corpus are added to all connective variants, complex forms, modifications and realizations, as well as to connective and non-connective usages and the whole lemmas.

The numbers reflect the total occurrences as well as intra-sentential (as opposed to inter-sentential) occurrences using the whole PDiT 2.0 data.

3.3 Example of a Lexicon Entry

The following is a shortened entry for a connective *proto* [therefore] (e.g. we shortened or deleted too long context examples and their English translations for better readability of the entry).

We may read the following information from the entry. E.g. 99% of all of its tokens in the PDiT are in a connective usage (i.e. its non-connective usage is very rare – cf. an example from the PDiT where *proto* [therefore] does not connect two discourse arguments but only two sentence elements: *Ještě ne na světové úrovni, a právě proto tak rozkošně živoucí.* [Not yet on the world level and exactly therefore so adorably lively.]). 28% within all of its connective usages is intra-sentential, which demonstrates the preference of this connective in inter-sentential discourse relations.

We may see that most preferably (in 98%), the connective signals a relation of reason-result (semantically, it expresses result). It appears in the second discourse argument and concerning its integration, the connective is not strictly bound to any position in the sentence (it may be used e.g. in the first as well as in the second position). 19% within the reason-result relation is formed by complex forms like *a proto* [and therefore] or *proto také* [therefore also] and 1% by modified forms like *právě proto* [exactly therefore].

Concerning other semantic types of discourse relations, the connective *proto* [therefore] expresses also pragmatic reason-result or equivalence; however, these relations are rather rare in this case.

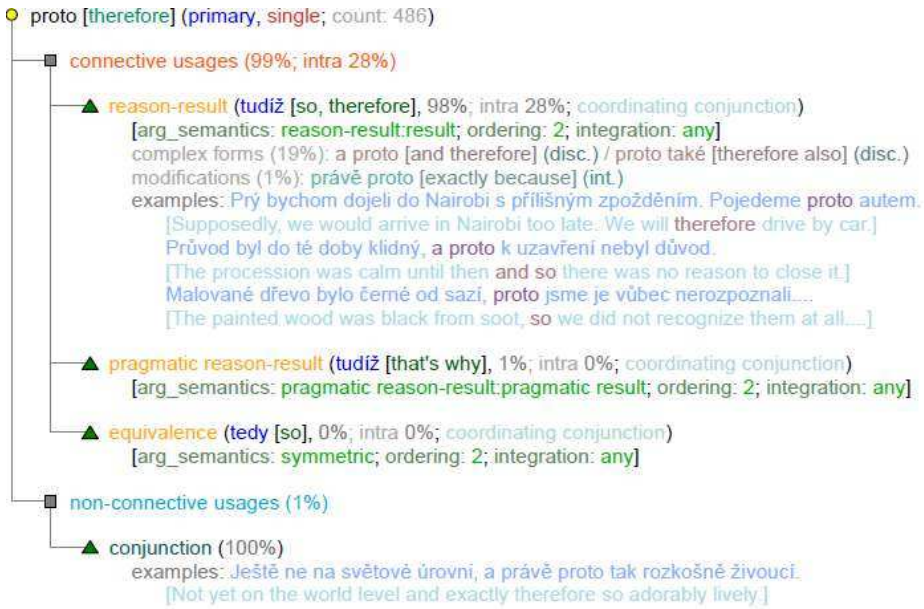

Fig 1. A shortened lexicon entry for the connective *proto* [therefore] in CzeDLex.

4 Conclusion

CzeDLex in its present version contains 205 lemmas (i.e. basic lemmas of primary and core words of secondary connectives) – all of them are manually annotated for modifications, complex forms, and variants. An additional manual annotation is provided to the most frequent ones, currently for primary connectives with at least 300 occurrences in the source corpus, and several most frequent secondary connectives.

Altogether, 19 lemmas have been fully manually processed, which covers more than two thirds of all discourse relations in the source corpus. Although the annotation of the rest of lemmas in CzeDLex is still in progress, we demonstrated that its first version offers valuable linguistic information already in its current form.

Acknowledgements

We acknowledge support from the Czech Science Foundation project no. GA17-06123S (*Anaphoricity in Connectives: Lexical Description and Bilingual Corpus Analysis*). This study has utilized the language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

1. Mírovský, J., Synková, P., Rysová, M., Poláková, L.: CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, No. 109, Copyright © Univerzita Karlova v Praze, Prague, Czech Republic, ISSN 0032-6585, 61–91 (2017).
2. Mírovský, J., Synková, P., Rysová, M., Poláková, L.: *CzeDLex 0.5*. Data/software, Charles University, Prague, Czech Republic, http://hdl.handle.net/11234/1-2538 (2017).
3. Mírovský, J., Synková, P., Rysová, M., Poláková, L.: Designing CzeDLex – A Lexicon of Czech Discourse Connectives. In: *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation*, pp. 449–457. Copyright © Kyung Hee University, Seoul, Korea, ISBN 978-89-6817-428-5 (2016).
4. Pajas, P., Štěpánek, J.: Recent Advances in a Feature-Rich Framework for Treebank Annotation. In: *The 22nd International Conference on Computational Linguistics – Proceedings of the Conference*, pp. 673–680. Copyright © The Coling 2008 Organizing Committee, Manchester, UK, ISBN 978-1-905593-45-3 (2008).
5. Poláková, L., Jínová, P., Zikánová, Š., Hajičová, E., Mírovský, J., Nedoluzhko, A., Rysová, M., Pavlíková, V., Zdeňková, J., Pergler, J., Ocelák, R.: *Prague Discourse Treebank 1.0*. Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://ufal.mff.cuni.cz/pdit/ (2012).
6. Poláková, L.: *Discourse Relations in Czech*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic, 197 pp. (2015).
7. Rysová, M., Rysová, K.: Secondary Connectives in the Prague Dependency Treebank. In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 291–299. Copyright © Uppsala University, Uppsala, Sweden, ISBN 978-91-637-8965-6 (2015).
8. Rysová, M., Rysová, K.: The Centre and Periphery of Discourse Connectives. In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pp. 452–459. Copyright © Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok, Thailand, ISBN 978-616-551-887-1 (2014).
9. Rysová, M., Synková, P., Mírovský, J., Hajičová, E., Nedoluzhko, A., Ocelák, R., Pergler, J., Poláková, L., Pavlíková, V., Zdeňková, J., Zikánová, Š.: *Prague Discourse Treebank 2.0*. Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://hdl.handle.net/11234/1-1905 (2016).
10. Rysová, M.: *Diskurzní konektory v češtině (Od centra k periferii)* [Discourse connectives in Czech (From Centre to Periphery)]. Ph.D. thesis, Charles University in Prague, Prague, Czechia, 268 pp. (2015).

11. Scheffler, T., Stede, M.: Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In: *Proceedings of LREC 2016*, pp. 1008–1013. European Language Resources Association, Paris, France (2016).
12. Stede, M., Umbach, C.: DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In: *Proceedings of Coling 1998*, pp. 1238–1242. Association for Computational Linguistics (1998).
13. Stede, M.: DiMLex: A lexical approach to discourse markers. In: *Exploring the Lexicon – Theory and Computation*. Alessandria (Italy): Edizioni dell'Orso (2002).
14. Synková, P., Rysová, M., Poláková, L., Mírovský, J.: Extracting a Lexicon of Discourse Connectives in Czech from an Annotated Corpus. Accepted for publication in: *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pp. 1–9. Copyright © Computing Society of the Philippines, Cebu, Philippines (2017).
15. Štěpánek, J., Pajas, P.: Querying Diverse Treebanks in a Uniform Way. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 1828–1835. Copyright © European Language Resources Association, Valletta, Malta, ISBN 2-9517408-6-7 (2010).
16. Zikánová, Š., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., Václ, J.: *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Copyright © ÚFAL, Praha, Czechia, ISBN 978-80-904571-8-8, 274 pp. (2015).