PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 117 OCTOBER 2021 5–26

Extending Coverage of a Lexicon of Discourse Connectives Using Annotation Projection

Jiří Mírovský, Pavlína Synková, Lucie Poláková

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

Abstract

We present a method for extending coverage of the Lexicon of Czech Discourse Connectives – CzeDLex – using annotation projection. We take advantage of two language resources: (i) the Penn Discourse Treebank 3.0 as a source of manually annotated discourse relations in English, and (ii) the Prague Czech–English Dependency Treebank 2.0 as a translation of the English texts to Czech and a link between tokens on the two language sides. Although CzeDLex was originally extracted from a large Czech corpus, the presented method resulted in an addition of a number of new connectives and new types of usages (discourse types) for already present entries in the lexicon. We classify and elaborate on reasons why the rest of automatically preselected candidates were excluded from the process, and give examples of actual new additions.

1. Introduction

A growing interest in text coherence-aware methods can be traced in many areas of natural language processing (NLP), including tasks such as machine translation (Xiong et al., 2019; Meyer and Webber, 2013), text generation (Kiddon et al., 2016), summarization (Zhang, 2011), information extraction, opinion mining (Turney and Littman, 2003), coherence evaluation (Rysová et al., 2016), or machine translation evaluation (Bojar et al., 2018). Many of these tasks incorporate discourse parsing in text pre-processing and, naturally, discourse parsing methods have come into focus of the discourse research community, including two CoNLL shared tasks (Xue et al., 2015, 2016).

Discourse parsing methods can strongly benefit from two types of language resources – text corpora manually annotated with discourse relations and lexicons of

 Corresponding author: mirovsky@ufal.mff.cuni.cz
Cite as: Jiří Mírovský, Pavlína Synková, Lucie Poláková. Extending Coverage of a Lexicon of Discourse Connectives Using Annotation Projection. The Prague Bulletin of Mathematical Linguistics No. 117, 2021, pp. 5–26.
doi: 10.14712/00326585.015.

discourse connectives. Discourse-annotated corpora date back to the Penn Discourse Treebank (PDTB; Miltsakaki et al., 2004) and the RST Discourse Treebank (Carlson et al., 2002), representing two dominant theoretical approaches to discourse coherence representation in text corpora – local (shallow) vs. global discourse models. Both approaches have been later followed for many languages.[1]

Electronic lexicons of discourse connectives – as an invaluable resource for both theoretical discourse research and automatic discourse processing – also date back almost two decades: an XML-based and machine readable DiMLex for German (Stede, 2002) and a more human-oriented DPDE, a dictionary of Spanish discourse markers (Briz et al., 2003). Since then, their number has been steadily rising, recently (since 2014) in connection with the COST Action TextLink, dedicated to discourse resources and representations: LexConn for French (Roze et al., 2012), LICO for Italian (Feltracco et al., 2016), CzeDLex for Czech (Mírovský et al., 2017), DiMLex-Eng for English (Das et al., 2018), LDM-PT for Portuguese (Mendes et al., 2018), and others. Most of these resources have been gradually incorporated in Connective-Lex (Stede et al., 2019), a multi-language database of discourse connectives currently covering 10 languages.[2]

Lexicons of discourse connectives gather and organize structured information about discourse connectives. Discourse connectives are words or phrases explicitly signalling discourse relations, i.e. semantico-pragmatic relations between two text spans (often called arguments). These relations can be either intra- or inter-sentential (i.e. they can occur within one sentence or between two or more sentences). Example 1 from the PDTB exhibits an intra-sentential discourse relation of discourse sense *Comparison.Concession* expressed with connective *though*.[3]

(1) **Though** Mrs. Thatcher has pulled through other crises, supporters wonder if her steely, autocratic ways are the right formula today. (PDTB)

We distinguish two types of discourse connectives, primary and secondary (Rysová and Rysová, 2014). Primary connectives form an almost closed set of mainly one-word expressions belonging mostly to conjunctions (*but*, *or*, *however*), particles (*only*, *too*) and adverbs (*later*, *previously*).[4] Secondary connectives belong to an open set of a broad range of expressions that are not yet fully stable or grammaticalized (*for these reasons*, *the main condition is*, *that is why*); they can be a part of the sentence syntac-

[1] A summarizing list of discourse-annotated corpora for different languages and within different frameworks can be found at `http://www.textlink.ii.metu.edu.tr/corpus-view`

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

[3] Discourse relations can be expressed by a connective (we call them explicit discourse relations), or understandable only from the context and the meaning of the arguments (we call them implicit).

[4] According to the traditional Czech word class categorization, particles form an autonomous category. In contrast to adverbs they do not participate in the sentence structure.

tic structure or even stand as a separate clause. Secondary connectives correspond roughly to alternative lexicalizations (AltLexes) in the PDTB terminology.

For using lexicons of discourse connectives in NLP tasks, it is crucial that the lexicons carry linguistic information not only about syntactic properties of the connectives, but most importantly also about their semantic properties, i.e. a list of discourse senses[5] the connective can express and the semantics of the discourse relation arguments (e.g., for the *reason-result* relation, which of the arguments represents the "reason" and which the "result"). Some of the lexicons were built with this principle from the start (LexConn, CzeDLex), some others were enriched with semantic information in their recent versions (DiMLex, Scheffler and Stede, 2016). All lexicons to be added to Connective-Lex are required to carry the semantic information.

Various strategies may be employed to build electronic lexicons of discourse connectives, depending on available resources – traditional printed lexicons may be consulted, discourse-annotated corpora may be used to extract lexicon data, various projection methods may be used to utilize existing discourse-related resources in another language, etc. However, in any case, building a lexicon with richly annotated entries requires a lot of (subsequent) manual work.

The Lexicon of Czech Discourse Connectives, CzeDLex (Mírovský et al., 2017), was originally extracted from a large Czech discourse-annotated treebank – the Prague Discourse Treebank 2.0 (PDiT; Rysová et al., 2016). The extraction from this 50-thousand-sentences corpus with more than 20 thousand annotated explicit discourse relations produced a lexicon with approx. 200 entries, which have been gradually manually edited since, leading to several published versions of the lexicon. CzeDLex 0.6, published in December 2019 (Synková et al., 2019), contained 204 entries, out of which 76 entries (covering more than 90% of the discourse relations annotated in PDiT) were fully manually checked and supplemented with additional linguistic information. It was the last version of CzeDLex containing solely entries originating in PDiT.

The present article elaborates on theoretical and practical aspects of the subsequent enrichment of the lexicon by exploiting the method of annotation projection and two additional resources – the Penn Discourse Treebank 3.0 (PDTB; Prasad et al., 2019) and the Prague Czech–English Dependency Treebank 2.0 (PCEDT; Hajič et al., 2012a).

Annotation projection is a well established and widely used method of automatic or partially automatic cost-effective linguistic annotation. The purpose of the projection is to induce annotation of a certain language phenomenon in a target language, using an already existing annotation of the phenomenon in a source language and parallel texts/corpora in the two languages.

[5] Throughout this article, the term (*discourse*) *senses* is used in compliance with the PDTB terminology when speaking about the senses/meanings of English, PDTB-style-annotated discourse relations, whereas the term *discourse* (*semantic*) *types* refers to the same notion in Czech annotations both in PDiT and in CzeDLex.

The method has been employed in various types of tasks, ranging from morphology to syntax and to semantics. To name just a few examples out of many, Yarowsky and Ngai (2001) used annotation projection for part-of-speech tagging and detection of noun phrases, with English as the source language and French and Chinese as the target languages. Hwa et al. (2005) trained dependency syntax parsers for Spanish and Chinese on data obtained by a projection of manual syntactic annotation in English. Padó and Lapata (2009) exploited possibilities of annotation projection from English to German on the task of semantic roles labeling.

Annotation projection is not unheard of either in the field of discourse annotation: Versley (2010) used annotation projection to induce detection of discourse connectives in German using English–German parallel texts and an automatic discourse parser on the English side. In 2017, Laali and Kosseim studied possibilities of projecting annotation of discourse relations from English to French, creating a discourse annotated French corpus of Europarl data. Sluyter-Gäthje et al. (2020) even used automatically translated texts of the PDTB (and annotation projection) to create a German discourse-annotated corpus, GermanPDTB.

The rest of the article is organized as follows. We describe our method and data in detail in Section 2 and analyze the results in Section 3. Section 4 concludes the article.

2. Data and Methodology

The possibility to use annotation projection to enrich the Lexicon of Czech Discourse Connectives, CzeDLex, with additional data extracted from another discourse-annotated treebank comes from a unique situation of having two key resources at our disposal, the PDTB 3.0 and the PCEDT 2.0:

PDTB 3.0: The Penn Discourse Treebank 3.0 (Prasad et al., 2019) is a corpus of English newspaper texts annotated manually with discourse relations. The texts consist of approx. 50 thousand sentences of the Wall Street Journal section of the Penn Treebank (Marcus et al., 1995) and the annotation contains approx. 40 thousand discourse relations of various kinds (including implicit relations and entity-based relations). For our purposes, we have used approx. 26 thousand relations explicitly expressed by a connective.

It is worth noting that PDiT (the original source corpus for CzeDLex)[6] and the PDTB are comparable in genre (journalism), size (50k sentences) and are similar also in the annotation scenario[7] and the extent of the annotated explicit discourse relations (21 thousand vs. 26 thousand).

[6] the Prague Discourse Treebank 2.0 (PDiT; Rysová et al., 2016)

[7] although there are differences in the sense hierarchies and e.g. implicit relations were not annotated in PDiT

PCEDT 2.0: The Prague Czech–English Dependency Treebank 2.0 (PCEDT; Hajič et al., 2012a, Hajič et al., 2012b) is a corpus of English–Czech parallel texts and their analysis on several layers of language description in the same annotation scenario as PDiT. Importantly, the English part of the PCEDT contains the same texts as the PDTB, i.e. the Wall Street Journal section of the Penn Treebank (PTB). The Czech part is based on human translations of the English texts to Czech, by design 1:1 sentence-aligned, with an additional automatic alignment on the word/node level on all annotation layers.

Methods used in the research described in the present article were implemented in the Prague Markup Langauge data format and application framework (PML; Pajas and Štěpánek, 2008), which is a primary format for PDiT, CzeDLex and the PCEDT. From a previous research, also the PDTB (mapped onto the PTB) was available in the PML format (Mírovský et al., 2016). The Prague Markup Language is an XML-based format and application framework designed for multi-layer linguistic annotations with available tools allowing for complex linguistic studies: tree editor TrEd[8] for browsing and editing PML data, `btred` for applying Perl scripts to the data and Prague Markup Language - Tree Query (PML-TQ; Pajas and Štěpánek, 2009) as a powerful, graphically oriented query system.[9]

The method for CzeDLex enrichment consisted of the following distinctive steps:

1. projection of the PDTB discourse annotation to the Czech part of the PCEDT (PCEDT-cz), see Sec. 2.1 below,
2. transformation of the PDTB discourse senses to the Prague taxonomy (Sec. 2.2),
3. extraction of czedlex-pcedt-cz, a raw PCEDT-cz-based lexicon of connectives (Sec. 2.3),
4. identification of connectives and discourse senses not present in CzeDLex, manual selection of the relevant ones (Sec. 2.4 and 3),
5. merging the selected new data into CzeDLex (Sec. 2.5),
6. manual fixes/annotation (an ongoing work).

We describe the individual steps from the technical point of view in Sections 2.1 – 2.5. Section 3 offers a detailed analysis of the manual selection in step 4 from a linguistic point of view. Step 6 represents an ongoing work, to be finished by the end of the year by a publication of a new version of CzeDLex.

[8] `https://ufal.mff.cuni.cz/tred/`

[9] See Mírovský et al. (2016) and Mírovský et al. (2014) for a demonstration how to search with the PML-TQ in the PDTB and PDiT, respectively.

2.1. Annotation Projection

The projection of the discourse annotation from the PDTB to the PCEDT-cz consisted of several sub-steps. The discourse annotation was first mapped from the raw texts to the Penn Treebank phrase structure trees using procedures and the framework described in Mírovský et al. (2016), newly adapted to the annotation scheme of version 3 of the PDTB. Among other things, the adaptation involved a computation of so called GORN addresses,[10] used to map text spans defined by character offsets in the raw texts to nodes in the trees of the Penn Treebank. After the mapping, all attributes of the discourse relations related to text spans became represented by minimal sets of nodes in the PTB, and the relations themselves became represented as links (arrows) between the sets of nodes corresponding to the discourse relation arguments.

Second, the discourse annotation was copied from the Penn Treebank phrase structure trees to the dependency trees of the tectogrammatical layer of the English part of the PCEDT (PCEDT-en), using 1:1 correspondence between terminal nodes of the Penn Treebank and nodes at the analytical (surface syntax) layer of the PCEDT-en (a-nodes), and then links from nodes on the tectogrammatical (deep syntax) layer of the PCEDT-en (t-nodes) to the a-nodes. The annotation obtained at this point was structurally close to the one in PDiT.

Finally, the discourse annotation was projected from English to Czech, i.e. from the PCEDT-en tectogrammatical trees to the PCEDT-cz tectogrammatical trees, using an automatic alignment of nodes on the corresponding t-layers. Errors originating from the automatic alignment form a large part of errors in the projected data and are discussed in Section 2.3 and also in Section 3.

2.2. Sense Taxonomy Transformation

The PDTB and PDiT use similar sets of senses/discourse relation types.[11] Table 1 shows the mapping of the PDTB 3.0 senses to PDiT discourse types used in the transformation.[12] The mapping is not entirely 1:1 – in cases when a single PDTB sense maps to two PDiT discourse types, the more frequent one was used (listed first in the table). Please note that only the sense *Expansion.Level-of-detail* distinguishes argument semantics in the table, as it maps to two different discourse types in Czech (*specifica-*

[10] which were a part of the published PDTB 2.0 data but are not a part of the PDTB 3.0 data

[11] The set of discourse types in PDiT was originally inspired by the Penn Discourse Treebank 2.0 sense hierarchy (Prasad et al., 2008).

[12] A similar transformation table is published in Mírovský and Poláková (2021) with a few mistakes: *Comparison.Contrast* was transformed to *opposition* (should be *confrontation*), *Comparison.Concession+SpeechAct* was transformed to *confrontation* (should be *pragmatic opposition*). The senses *Contingency.Negative-cause* and *Expansion.Manner* do not in fact have a counterpart in the Czech taxonomy (these types of relations were not annotated in PDiT) but the table erroneously mapped the senses to discourse types *reason–result* and *explication*, respectively. Table 1 in the present article fixes the errors.

PDTB 3 sense	PDiT 2 discourse type
Comparison.Concession	concession, opposition
Comparison.Concession+SpeechAct	pragmatic opposition
Comparison.Contrast	confrontation
Comparison.Similarity	conjunction
Contingency.Cause	reason–result
Contingency.Cause+Belief	explication
Contingency.Cause+SpeechAct	pragmatic reason–result
Contingency.Condition	condition
Contingency.Condition+SpeechAct	pragmatic condition
Contingency.Negative-cause	-
Contingency.Negative-condition	condition
Contingency.Purpose	purpose
Expansion.Conjunction	conjunction, gradation
Expansion.Disjunction	disjunctive alternative, conjunctive alt.
Expansion.Equivalence	equivalence
Expansion.Exception	restrictive opposition
Expansion.Instantiation	instantiation
Expansion.Level-of-detail.Arg1-as-detail	generalization
Expansion.Level-of-detail.Arg2-as-detail	specification
Expansion.Manner	-
Expansion.Substitution	correction
Temporal.Asynchronous	precedence–succession
Temporal.Synchronous	synchrony

Table 1. The PDTB – PDiT sense transformation table.

tion and *generalization*). Argument semantics of other asymmetric senses projected to the PDiT taxonomy is captured in the direction of the discourse relation (which is represented as a link and depicted by an arrow in tectogrammatical trees).[13]

2.3. Extraction of a Raw Lexicon

A raw version of CzeDLex was originally extracted from PDiT. The extraction script used a flat list of connectives occurring in the annotated data, manually pre-grouped in the sense of a connective and its variants, modifications and complex forms. The script took the flat list of grouped connectives, went through the annotated discourse data and integrated information from each discourse relation into the raw lexicon, gradually creating entries for individual connectives and their possible discourse types.

[13] Argument semantics specifies roles of two arguments of an asymmetric discourse relation – e.g., for discourse type *reason–result*, it specifies which of the arguments is the *reason* and which one is the *result*.

Additional information was also collected by the script, such as argument semantics, numbers of occurrences, corpus examples, examples of non-connective usages and others. The extraction process was described in detail in Synková et al. (2017). Manual annotation and corrections then started on these automatically extracted data (the whole process was summarized in Mírovský et al., 2017).

Similar scripts were now used to automatically extract raw lexicon data from the PDTB discourse annotation projected to the PCEDT-cz (we only used relations of types Explicit, AltLex a AltLexC, i.e. relations originally expressed by a connective, with a non-empty counterpart in the word alignment). The flat list of connectives appearing in the PCEDT-cz contained almost 3 thousand entries, most of them representing errors in the word alignment between the English and Czech parts of the PCEDT. We first cut off all single occurrences (over 2 500 entries). From the remaining slightly over 500 entries, the most obvious word alignment errors were deleted, the rest of entries were pre-grouped in the sense of modifications etc. The resulting grouped flat list was used to automatically extract a raw lexicon czedlex-pcedt-cz from the PCEDT-cz, containing over 200 entries (connectives) along with their variants, possible discourse types, complex forms, modifications, examples (original English and Czech translations), corpus counts etc.

2.4. Automatic Pre-Selection and Manual Selection

The extracted czedlex-pcedt-cz was automatically compared with the current version of CzeDLex to mark connectives not appearing in CzeDLex and – for connectives already present in CzeDLex – to mark discourse types not appearing at the respective entries in CzeDLex. This marking produced a list of 92 potential candidates for new connective entries and further 250 potential new discourse types to be added to existing entries.

These candidates were subsequently inspected by two experienced annotators who were asked to mark each candidate (a whole entry or a discourse type) with one of three options meaning `USE`, `POSSIBLY USE` and `DO NOT USE`. The annotators considered the automatically collected examples and (if needed) their broader textual context both in Czech and in original English, for complex cases they entered comments and discussed their choices. This process significantly narrowed the selection of candidates and is analyzed in detail in Section 3.

Only candidates that were marked at least by one of the annotators as `USE` or by both as `POSSIBLY USE` (in total, 25 new whole entries and 17 new discourse types for already existing entries) were then actually selected for an inclusion into CzeDLex.

2.5. Merging

The selected 25 connectives and additional 17 discourse types were merged to the current version of CzeDLex, being still subjects to a later detailed manual inspection

and annotation (and even possible eventual deletion) just like any previous CzeDLex entry/discourse type. A new attribute `source` marking external source was added to the data scheme and filled with value PCEDT for the new data. In the graphical environment, the external source is clearly visible, distinguishing the new data from the original ones. New discourse types of already existing connectives are sorted at the end of possible discourse types of a connective and their counts in the source corpus are not added to the overall count of the connective (otherwise they would disrupt percentages of various discourse types for the connective).

3. Analysis of the Projected Data

This section of the article addresses the process of manual inspection and evaluation of the extracted and pre-selected lexicon data by the annotators. We describe a set of connective and discourse type candidates (out of the automatically pre-selected 92 and 250, resp.) that were in the end not included in CzeDLex and categorize reasons for their exclusion (Section 3.1). Then we present the set of included new connectives and discourse types (Section 3.2). The discussed categories are accompanied by corpus examples, i.e. Czech translations from the Czech part of the PCEDT and the English PDTB originals. Connectives in the examples are highlighted in bold.

3.1. Candidates Not Included in CzeDLex

The reasons for not including some of the pre-selected candidates to CzeDLex can be divided into three main groups; we address them below in detail:

1. differences in the annotation schemes and strategies (Sec. 3.1.1),
2. issues coming from the translation (Sec. 3.1.2),
3. errors originating in the projection process (Sec. 3.1.3).

3.1.1. Differences in the annotation schemes

The issues arising from the differences in the annotation schemes for English (the PDTB approach) and Czech discourse relations (the PDiT approach) included namely the following: differences in distinguishing across discourse senses/types and in definitions of individual senses/discourse types (different annotation guidelines for senses/discourse types with the same label), and differences in the evaluation of individual expressions with respect to actually fulfilling the function of a connective (in a given context). Altogether, these reasons account for approx. 40% of the excluded candidates; the majority of these cases represent differences in the semantic label taxonomies and annotation strategies.

The semantic taxonomies of the PDTB 3.0 and PDiT 2.0 differ first in the presence of senses *Expansion.Manner* and *Comparison.Similarity*, which the PDiT approach con-

siders to be rather a part of the syntactic analysis[14] or a specific case of *conjunction* (*Comparison.Similarity*), and second, in the absence of the relation of *gradation* (a part of *Expansion.Conjunction* in the PDTB), *explication* (a part of *Contingency.Cause* and *Contingency.Cause+Belief*) and *conjunctive alternative* (a part of *Expansion.Disjunction*). Mismatches caused by these differences are complemented by cases where our annotators did not agree with the PDTB interpretation of the given relation.

Example 2 shows a PDTB relation *Expansion.Manner.Arg2-as-manner* that is not considered to be a discourse relation in PDiT; Example 3 shows a PDTB relation with the same label that would be interpreted as *conjunction* in the PDiT approach, and Example 4 illustrates a context in which we do not agree with the PDTB interpretation – this context in our opinion contains an *Expansion.Equivalence* relation, not *Expansion.Conjunction*.

(2) Potom jsem si všimla, že se auto pohybuje nahoru a dolů, **jako kdyby** na něm někdo skákal.

[Then I noticed the car was bouncing up and down **as if** someone were jumping on it.]

(3) Ale firma Honda letos model Accord zrenovovala **a** udělala z něj vůz střední velikosti.

[But this year, Honda has revamped the Accord **and** made it a midsized car.]

(4) Podle Cathcartových slov to bude ve společnosti Kidder v nadcházejících letech „hučet jako v úle". **Neboli**, jak říká Carpenter opírající se o své zkušenosti z konzultantské firmy: „Teď jsme připraveni jednat."

[In coming years, Mr. Cathcart says, Kidder is "gonna hum." **Or**, as Mr. Carpenter, again drawing on his consulting-firm background, puts it: "We're ready to implement at this point."]

Differences in classifying certain words (tokens) as connectives are a reason for excluding, e.g., a comma from the set of new connectives, since a comma has in our opinion too many other functions to be used as a reliable signal of a discourse relation,[15] the adverb *nyní* [*now*] was excluded for being considered a part of a bridging relation and a semantic constituent of the sentence rather than a connective. A specific case is represented by non-finite verb structures where the verb form itself is consid-

[14] The syntactic label *Expansion.Manner* in the underlying Czech syntactic annotation was not assessed to hold analogically in discourse annotation, as the possibilities of its expression in an inter-sentential setting seemed quite restricted in Czech and most similar cases were judged quite satisfactorily as *specification*.

[15] In the PDiT approach, only a colon, a semicolon and a dash are considered to be connectives.

ered both a part of an argument and a connective in the PDTB – cf. the verb *zanechat* [*leaving*] in Example 5.

In the PDiT approach, such (notional) verbs represent the core of a proposition, they are a constitutive part of the argument and thus do not exhibit the main feature of a connective, i.e. being an operator connecting two spans of a text.[16]

(5) Použití herbicidů by vybilo plodné rostliny a **zanechalo** velké pole rostlin se samčí sterilitou, které mohou být opylovány pro získání křížených semen.

[The application of herbicide would kill off the male-fertile plants, **leaving** a large field of male-sterile plants that can be cross-pollinated to produce hybrid seed.]

3.1.2. Issues coming from the translation

Although the human translators of the PCEDT texts from English to Czech were instructed to translate as literally as possible (but fluently), differences originating in the translation are the most common cause for excluding candidates from the CzeDLex enrichment (they account for approx. 50%). During the analysis of the projected data, three main types of differences caused by translation differences were detected.

The most common type was a choice of a more specific Czech connective for a less specific English one, e.g. *as* with a temporal annotation in Example 6 was translated as *jelikož* [*because*]. In English, *as* is a highly polyfunctional expression (according to the PDTB annotation, it can signal relations from all four major classes of discourse senses). In Czech, and similarly in other languages, a translation of expressions such as *as* necessarily implies a disambiguation among the possible interpretations of the original word.[17] The Czech translation equivalent in Example 6 *jelikož* signals the meaning of *reason–result*, it does not have a temporal meaning. Thus, the new temporal meaning of the connective *jelikož* coming from the PDTB projection had to be excluded from the CzeDLex enrichment.

(6) V Londýně při nestálém obchodování uzavřely akcie níže, **jelikož** začínající zotavení bylo zeslabeno obchodními výsledky USA, které jsou horší, než se čekalo.

[In London, stocks closed lower in volatile trading **as** an opening rally was obliterated by worse-than-expected U.S. trade figures.]

[16] The issue of verb forms representing a connective is a more general one. Surely some verbs have some inner connectivity feature (*imply, cause, mean, contradict, follow...*) but whether to assess them as connectives or as arguments (propositions), or, where to set the border, is a theoretical question for discussion. At present, the Czech annotations and the CzeDLex do not include verbs as connective entries.

[17] Disambiguation by translation in general is a well-known topic in translation studies, but also a separate topic in discourse research: disambiguation of (functions of) connectives by their translation, e.g. Meyer (2011), Cartoni et al. (2013).

The second type of "translation reasons" for excluding candidates for CzeDLex enrichment were different properties of English–Czech counterparts both at the word and the higher construction level. A difference at the word level is illustrated in Example 7. English *indeed* can be in many contexts translated correctly as *opravdu*,[18] but it cannot stand separately at the very beginning of a sentence. Being a constituent within the sentence, it loses its connectivity and becomes a modal particle – thus in the present context, it would be more appropriate to leave it out or to choose a non-literal translation. The connectivity of English *indeed* is beyond all doubt.

(7) Jeho organická architektura odrážela citlivý vztah k životnímu prostředí již desítky let předtím, než se toto téma stalo populární mezi „rádoby aktivisty". Wright **opravdu** celý svůj život tvrdil, že nejvíce se toho naučil studiem přírody.

[Wright's organic architecture demonstrated a keen sensitivity to the environment decades before it became fashionable among "la-la activists". **Indeed**, Wright said all his life that the greatest lessons he learned were derived from the study of nature.]

As for differences at the higher construction level, the most common case was the translation of an English non-finite verb structure by a Czech clause. The English structure *by taking...* in Example 8 could be translated by a Czech non-finite structure but the sentence would sound unnatural. The chosen translation (*když nastoupil...* [*when he took over...*]) sounds natural but does not preserve the meaning of the English structure – it is no longer *Contingency.Purpose* or *Expansion.Manner* (as annotated in the PDTB), but *synchronous* or *reason–result*.

(8) Fromstein upevnil svou kontrolu v dubnu, **když** nastoupil po Berrym na místo předsedy představenstva.

[Mr. Fromstein solidified his control in April **by** taking over from Mr. Berry as chairman.]

Apart from the translation by a dependent clause, another option is to translate an English non-finite verb structure by a Czech verbal noun – cf. Example 9 where the structure *for loading...* was translated as *k naložení...* where *k naložení* is a prepositional phrase with the noun in dative.

(9) Sovětské nákupy jsou tak masivní, že vývozci mají potíže sehnat dostatek říčních člunů a vlaků, aby dopravili právě sklizenou středozápadní úrodu do přístavů **k** naložení na sovětské lodě.

[The Soviet purchases are so massive that exporters are struggling to find

[18] or, maybe, more precisely as *vskutku*

enough river barges and trains to move the recently harvested Midwest crop to ports **for** loading onto Soviet ships.]

However, when translated by nouns, text spans forming a discourse argument in the PDTB do not represent an argument in the PDiT approach, as a PDiT argument requires a finite verb as its core[19] and these cases were thus excluded from the candidates to CzeDLex enrichment.

Finally, some candidates had to be excluded due to an inadequate translation. In Example 10, the English connective *while* is translated by Czech *místo aby* [*instead of*], which substantially modifies the sentence meaning. The connective–sense pair (in this context *místo–concession*) is thus unusable.

(10) **Místo aby** sliby ohledně velkých zisků rozezněly zvony na poplach, tak toho obvykle nedocílí, částečně proto, že povídačky o tom, jak se dá rychle zbohatnout, se staly pevnou součástí amerického folklóru.

[**While** the promises of big profits ought to set off warning bells, they often don't, in part because get-rich-quick tales have become embedded in American folklore.]

3.1.3. Projection errors

The least common reason for excluding candidates were projection errors, these cases accounted for approx. 10%. Most of these cases were less obvious errors caused by the automatic word alignment in the PCEDT that had not been detected before in the projection scenario (see Section 2.3). One context with such an error is in Example 11 – although in some other contexts the alignment correctly matched *although* and *i když* as counterparts, in this sentence it picked just the word *i* as the counterpart to *although* (producing a nonsensical connective–sense pair).

(11) **I když** se sledovanost v dobách převratných novinek prudce zvýší, v době zklidnění upadá.

[**Although** viewership soars when big news breaks, it ebbs during periods of calm.]

Another example of an error in the word alignment originates in different properties of English and Czech at the structural level. In Example 12 the English connective *yet* is translated by the Czech counterpart *však*, which has a substantially different word order position in this context. The word alignment wrongly picked the more conve-

[19] This was a practical annotation decision; we are aware of the fact that also non-finite verb structures and deverbative nouns can represent an argument of a discourse relation.

niently positioned adverb *dosud* in the temporal meaning of *so far*[20] as the translation of the connective *yet*, thus (wrongly) associating the relation of *concession* with the temporal connective.

(12) Koloběh bohužel nepřichází ve vlnách, ale v sestupné spirále. Důkazem toho, že dosud nejsme úplně na dně, je **však** to, že si ještě navzájem nepomáháme.

[Sadly, the cycle appears not as waves but as a downward spiral. **Yet** the evidence that we have not hit bottom is found in the fact that we are not yet helping ourselves.]

3.2. Candidates Included in CzeDLex

The list of candidates included to CzeDLex after the manual assessment comprises 25 new whole entries (connectives) and 17 new discourse types for already existing entries. As we wanted to eliminate a possible influence of "translationese" in connective translations, the frequency of new lemmas was checked in a large representative corpus of Czech (Křen et al., 2019) on original Czech texts and for some cases, even professional translators were consulted.

3.2.1. Whole new entries

Primary connectives newly added to CzeDLex as whole new entries are quite rare. They are mostly single-word adverbs and they represent less frequent alternatives to some more common primary connectives. Most of them did not occur in the original PDiT corpus and their existence and connective function was first documented in the PDTB translation, e.g. *kupříkladu* (more commonly *například* [*for example*]), *obdobně* (more commonly *podobně* [*similarly*], see Example 13), *taktéž* (*též, také* [*also, too*]). These new connectives and their more common counterparts are synonyms and have identical discourse functions, although there might be a slight difference in register: the new lemmas appear more formal than the more frequent ones.

(13) Například 88% čtenářů tohoto listu vlastní akcie (což je o něco méně než 91% v obdobném průzkumu loni). [Ale jen 17.5 % uvedlo, že mají na akciovém trhu více než polovinu svých peněz.] **Obdobně** 57% respondentů vlastní podíl v nějakém investičním fondu peněžního trhu a 33% vlastní komunální obligace.

[For example, about 88% of Journal readers owned stock (down slightly from 91% in a similar poll last year). [But only 17.5% said they had more than half

[20] which was actually introduced by the translator, the modification *so far* is not present in the original clause *we have not hit bottom*.

their money in the stock market.] **Similarly**, 57% of respondents own shares in a money-market mutual fund, and 33% own municipal bonds.] [21]

Secondary connectives are much more frequent as new additions to CzeDLex. In these cases, the core word is mostly a preposition (*namísto* [*instead*], see Example 14) or a noun (*doba* [*time, point*], Examples 15 and 16) and the whole connective is a phrase, the exact formulation of which is largely dependent on the chosen way and syntactic possibilities of the translation. In Example 14,[22] the connective is a phrase with a demonstrative pronoun *toho* (lit. *instead of that*), but in other contexts it can read: *namísto toho, aby* or *namísto*(,) *aby* (lit. *instead of that/the fact that*) – introducing a dependent clause of substitution in Czech. The order of the arguments can switch for different realizations. This fact is reflected in the lexicon by numbering in the `discourse type` attribute (e.g. *correction-1*, *correction-2*, etc.) and by distinguishing the numbered types by further attributes, according to the syntactic structure[23] that underlies the relation. For newly added secondary connectives, these distinctions are subject to finer manual work.

(14) Pro nové akcie nebyla dosud stanovena žádná cena. **Namísto toho** ponechají společnosti na trhu, ať rozhodne.

[No price for the new shares has been set. **Instead**, the companies will leave it up to the marketplace to decide.]

Connective candidates with the core word *doba* [*time, point*], or *okamžik* [*moment*], were added to CzeDLex, as they signal a non-negligible number of temporal relations. Originally, CzeDLex did not include such temporal nouns. They are various phrases containing the core expression, as the pre-processed entry was automatically merged from all instances of this expression within any word chain with a connective function.

For the connective with the core word *doba*, both temporal discourse types have been projected (*synchrony*, *asynchrony*). For the discourse type of *synchrony*, these phrases include four diverse translations[24] of the original English connectives *at the same time*, *at that time*, *at the time*, *at that point*, see Example 15. The same amount of translations[25] is documented for the discourse type of *asynchrony* and the original connectives *ever since*, *by then*, *until then*, *until*, see Example 16. This example nicely

[21] In this context, the original PDTB sense is *Similarity*, which was transferred to *conjunction* in the Czech taxonomy.

[22] Like the mentioned primary connectives, *namísto* is also a connective with a more common alternative (*místo* [*instead*]) and a possible slight shift in register towards formality.

[23] marked by attributes `schema` and `realizations`

[24] *ve stejné době*, *v té době*, *tou dobou*, *v té samé době*. The last one of them, in our opinion, is in Czech a rather awkward calque of the English *at the same time*.

[25] *od té doby, co*; *v té době*; *do té doby*; *do doby, než*

demonstrates two things: first, the many-to-many translation possibilities of connectives and the effect of the projection in bringing them together, and, the lexicographic challenge in the attempt to systematically capture secondary connectives. For the final record in CzeDLex, this entry will also need a significant manual detailing (schemas, realizations etc.).

(15) Posuďte zkušenosti Satoka Kitady, třicetiletého návrháře interiérů vozů, který nastoupil do firmy Nissan v roce 1982. **V té době** byly úkoly přidělovány striktně podle služebního věku.

[Consider the experience of Satoko Kitada, a 30-year-old designer of vehicle interiors who joined Nissan in 1982. **At that time**, tasks were assigned strictly on the basis of seniority.]

(16) **Od té doby, co** bylo spojení s cholesterolem odhaleno, začali Američané přidávat psylium do obilovin ke snídani.

[**Ever since** the link to cholesterol was disclosed, Americans have begun scarfing up psyllium in their breakfast cereals.]

Apart from temporal nouns as core words of new secondary connectives, causative or argumentative nouns as core words extended the list of CzeDLex entries, the original CzeDLex was more reserved in this respect. The newly added noun-based entries include core words *srovnání* [*in/by comparison*], *předpoklad* [*assuming*, *assuming that*, *providing*], *kontext* [*in that context*], *kontrast* [*in contrast*], *následek* [*as a result*, *to result in*], *základ* [*by*, *assuming*, lit. *based on*], *známka* [*indication*].

Other newly added candidates include adverbs *dříve* (*než*) [*before*, *previously*, *until*], *původně* [*originally*, *previously*], *skutečně* [*indeed*, *in fact*], focussing particles *především* [*in particular*, *especially*], *zejména* [*in particular*], *zvlášť* [*separately*], multiword connectives *než aby* [*rather than*] and phrases with prepositions *během* [*while*, *as*], *kromě* (*jiného*) [*among* (*other things*)].[26]

3.2.2. New discourse types in existing entries

The new 17 discourse types from the projection enriched 12 different already existing CzeDLex entries – five of the connectives were provided with even two new discourse types. An example entry is the connective *jak* [*as*, *when*]. This expression had originally documented four PDiT discourse types (*synchrony*, *asynchrony*, *reason–result*, *condition*) and a number of non-connective usages as well (e.g. *how*). The projection from the PDTB revealed two more connective usages, the discourse types of *concession*, see Example 17, and *instantiation*, Example 18. These usages may not be very frequent but

[26] Some of the newly added connectives are also present in the original PDiT corpus but – for various reasons, incl. simple omission – they were not annotated as connectives before.

they are fully acceptable. According to the sense – discourse type label mapping, the two added relations are identical in both corpora.

(17) Řekl, že sleduje údaje o peněžních zásobách, avšak nepřisuzuje jim prvořadou důležitost, **jak** navrhují někteří soukromí a vládní ekonomové.

[He said he monitors the money-supply figures, but doesn't give them paramount importance, **as** some private and government economists have suggested.]

(18) Taková situace může způsobit spoušť, **jak** ukázal mimořádný případ, který v Chicagské obchodní komoře nastal toto léto v termínovém obchodu se sójovými boby.

[Such a situation can wreak havoc, **as** was shown by the emergency that developed in soybean futures trading this summer on the Chicago Board of Trade.]

As for mappings to a different PDiT discourse type, Example 19 with the *although* connective represents a PDTB *Expansion.Exception*. The meaning of exception is included in the Czech label *restrictive opposition*, which covers both exceptions and "milder" restrictive contrasts.

(19) Všichni jsme tu v pořádku, **ačkoliv** Mame byla nesmírně vystrašená.

[We are all fine here, **although** Mame was extremely freaked.] [27]

4. Conclusion

Coverage (or completeness) of any lexicon is one of its key aspects. We have presented a method for extending coverage of the Lexicon of Czech Discourse Connectives, CzeDLex, using data obtained via annotation projection from a discourse-annotated corpus in English. The process resulted in an inclusion of 25 new full entries and 17 new discourse types for already existing entries.

Translated texts are of a different nature in comparison with texts written originally in a given language. It may be a question of discussion whether it is desirable to expand an original-text-based language resource (a lexicon) by data coming from exploiting translated texts. On the other hand, from the practical point of view, NLP applications using the lexicon should be able to process not only perfect Czech texts but also translated texts and maybe even awkward translations. To address this issue in CzeDLex, we have employed two measures. First, all data originating from English translations are clearly marked as such, and new discourse types for previ-

[27] The translation of the English *although* to Czech in this context was discussed with professional translators. It appears that, while grammatically correct, there are at least two much better translation options to make the sentence sound more natural in Czech.

ously present connectives are kept separately (at the end of the list); it applies also to corpus counts of these discourse types. Second, whenever in doubt, the annotators consulted a large corpus of Czech texts to check expressions that might sound unnatural to Czech native speakers.

All connectives and discourse types added to CzeDLex are subjects to a subsequent detailed check and annotation just like any previous entry. It is therefore possible that before the final publication, some of the new additions will be deleted, merged with another entry or otherwise modified, and in any case supplemented with additional linguistic annotation. The new version of CzeDLex is planned for publication by the end of 2021 in the LINDAT/CLARIAH-CZ repository[28] under the Creative Commons license.

Acknowledgements

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

Bibliography

Bojar, Ondřej, Jiří Mírovskỳ, Kateřina Rysová, and Magdaléna Rysová. Evald Reference-less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 541–545, 2018. doi: 10.18653/v1/W18-6432.

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

Carlson, Lynn, Mary Ellen Okurowski, Daniel Marcu, et al. *RST Discourse Treebank*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, 2002.

Cartoni, Bruno, S. Zufferey, and T. Meyer. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue Discourse*, 4:65–86, 2013. doi: 10.5087/dad.2013.204.

Das, Debopam, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. Constructing a Lexicon of English Discourse Connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365, 2018. doi: 10.18653/v1/W18-5042.

Feltracco, Anna, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. LICO: A Lexicon of Italian Connectives. *CLiC it*, page 141, 2016. doi: 10.4000/books.aaccademia.1770.

Hajič, Jan, Eva Hajicová, Jarmila Panevová, Petr Sgall, Ondrej Bojar, Silvie Cinková, Eva Fucíková, Marie Mikulová, Petr Pajas, Jan Popelka, et al. Announcing Prague Czech–English Dependency Treebank 2.0. In *LREC*, pages 3153–3160, 2012a.

[28] `https://lindat.cz`

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. *Prague Czech-English Dependency Treebank 2.0*. Data/Software, Linguistic Data Consortium, 2012b. University of Pennsylvania, Philadelphia. LDC2012T08.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural language engineering*, 11(3): 311–326, 2005. doi: 10.1017/S1351324905003840.

Kiddon, Chloé, Luke Zettlemoyer, and Yejin Choi. Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, 2016. doi: 10.18653/v1/D16-1032.

Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, et al. Korpus SYN, verze 8 z 12. 12. 2019, 2019.

Laali, Majid and Leila Kosseim. Improving Discourse Relation Projection to Build Discourse Annotated Corpora. *arXiv preprint arXiv:1707.06357*, 2017. doi: 10.26615/978-954-452-049-6_054.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. *Treebank-2*. Data/Software, Linguistic Data Consortium, 1995. University of Pennsylvania, Philadelphia. LDC95T7.

Mendes, Amália, Iria del Rio, Manfred Stede, and Felix Dombek. A Lexicon of Discourse Markers for Portuguese–LDM-PT. In *11th International Conference on Language Resources and Evaluation*, pages 4379–4384, 2018.

Meyer, Thomas. Disambiguating temporal-contrastive connectives for machine translation. In *Proceedings of the ACL 2011 Student Session*, pages 46–51, Portland, OR, USA, June 2011. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P11-3009`.

Meyer, Thomas and Bonnie Webber. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, 2013.

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/618.pdf`.

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, 2016. European Language Resources Association.

Mírovský, Jiří, Pavlína Synková, Magdaléna Rysová, and Lucie Poláková. CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, (109):61–91, 2017. ISSN 0032-6585.

Mírovský, Jiří and Lucie Poláková. Sense Prediction for Explicit Discourse Relations with BERT. In Yang, Xin-She, Simon Sherratt, Nilanjan Dey, and Amit Joshi, editors, *Proceedings of Sixth International Congress on Information and Communication Technology* (*ICICT*), volume 216 of *Lecture Notes in Networks and Systems*, pages 835–842, Singapore, 2021. International Congress and Excellence Awards, Springer.

Padó, Sebastian and Mirella Lapata. Cross-Lingual Annotation Projection for Semantic Roles. *Journal of Artificial Intelligence Research*, 36:307–340, 2009. doi: 10.1613/jair.2863.

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. doi: 10.3115/1599081.1599166. URL `https://www.aclweb.org/anthology/C08-1085.pdf`.

Pajas, Petr and Jan Štěpánek. System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL–IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, 2009. Association for Computational Linguistics. doi: 10.3115/1667872.1667881. URL `https://www.aclweb.org/anthology/P09-4009.pdf`.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC'08*), pages 2961–2968, Marrakech, 2008. European Language Resources Association. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.9566&rep=rep1&type=pdf`.

Prasad, Rashmi, Bonnie Webber, Alan Lee, and Aravind Joshi. *Penn Discourse Treebank Version 3.0*. Data/Software, Linguistic Data Consortium, 2019. URL `https://catalog.ldc.upenn.edu/LDC2019T05`. University of Pennsylvania, Philadelphia. LDC2019T05.

Roze, Charlotte, Laurence Danlos, and Philippe Muller. LEXCONN: A French Lexicon of Discourse Connectives. *Discours. Revue de linguistique, psycholinguistique et informatique*, (10), 2012. doi: 10.4000/discours.8645.

Rysová, Kateřina, Magdaléna Rysová, and Jiří Mírovskỳ. Automatic Evaluation of Surface Coherence in L2 Texts in Czech. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing* (*ROCLING 2016*), pages 214–228, 2016.

Rysová, Magdaléna and Kateřina Rysová. The Centre and Periphery of Discourse Connectives. In *Proceedings of Pacific Asia Conference on Language, Information and Computing*, pages 452–459, Bangkok, 2014. Department of Linguistics, Faculty of Arts, Chulalongkorn University. URL `https://www.aclweb.org/anthology/Y14-1052.pdf`.

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. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2016. URL `http://hdl.handle.net/11234/1-1905`.

Scheffler, Tatjana and Manfred Stede. Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 1008–1013, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L16-1160`.

Sluyter-Gäthje, Henny, Peter Bourgonje, and Manfred Stede. Shallow Discourse Parsing for Under-Resourced Languages: Combining Machine Translation and Annotation Projection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1044–1050, 2020.

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

Stede, Manfred, Tatjana Scheffler, and Amália Mendes. Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24), 2019. doi: 10.4000/discours.10098.

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 232–240, Cebu, Philippines, 2017. University of the Philippines Cebu. ISBN 978-89-6817-428-5.

Synková, Pavlína, Lucie Poláková, Jiří Mírovský, and Magdaléna Rysová. *CzeDLex 0.6*. Data/Software, Charles University, ÚFAL MFF UK, Prague, Czech Republic, http://hdl.handle.net/11234/1-3074, 2019.

Turney, Peter D and Michael L Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003. doi: 10.1145/944012.944013.

Versley, Yannick. Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82, 2010.

Xiong, Hao, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling Coherence for Discourse Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345, 2019. doi: 10.1609/aaai.v33i01.33017338.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 1–16, 2015. doi: 10.18653/v1/K15-2001.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, 2016. doi: 10.18653/v1/K16-2001.

Yarowsky, David and Grace Ngai. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001. doi: 10.3115/1073336.1073362.

Zhang, Renxian. Sentence Ordering Driven by Local and Global Coherence for Summary Generation. In *Proceedings of the ACL 2011 Student Session*, pages 6–11, 2011.

Address for correspondence:
Jiří Mírovský
mirovsky@ufal.mff.cuni.cz
Intitute of Formal and Applied Linguistics
Malostranské nám. 25
118 00 Prague 1
Czech Republic