Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1089–1096
Marseille, 11–16 May 2020

GeCzLex: Lexicon of Czech and German Anaphoric Connectives

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

Abstract

We introduce the first version of GeCzLex, an online electronic resource for translation equivalents of Czech and German discourse connectives. The lexicon is one of the outcomes of the research on anaphoricity and long-distance relations in discourse, it contains at present anaphoric connectives (ACs) for Czech and German connectives, and further their possible translations documented in bilingual parallel corpora (not necessarily anaphoric). As a basis, we use two existing monolingual lexicons of connectives: the Lexicon of Czech Discourse Connectives (CzeDLex) and the Lexicon of Discourse Markers (DiMLex) for German, interlink their relevant entries via semantic annotation of the connectives (according to the PDTB 3 sense taxonomy) and statistical information of translation possibilities from the Czech and German parallel data of the InterCorp project. The lexicon is, as far as we know, the first bilingual inventory of connectives with linkage on the level of individual entries, and a first attempt to systematically describe devices engaged in long-distance, non-local discourse coherence. The lexicon is freely available under the Creative Commons License.

Keywords: discourse connectives, bilingual lexicon, anaphoricity, Czech, German

1. Introduction

Recent years witnessed a boom in the development of electronic resources describing discourse relational devices (DRDs), or, more specifically, discourse markers and connectives in different languages (e.g. Feltracco et al. (2016), Mendes and Lejeune (2016), Das et al. (2018)). Efforts in this area were largely supported by the TextLink initiative,[1] which brought together discourse-oriented researchers from across Europe and some other countries. During the last two years, the first ones of these monolingual connective lexicons have been interlinked, and those developed later on then incrementally added, resulting in Connective-Lex, a multilingual connective database (Stede et al., 2019). Currently, Connective-Lex gathers lexicons for nine languages,[2] with their entries linked by the semantic taxonomy adopted by all the monolingual lexicons, the Penn Discourse Treebank 3 tagset (Webber et al., 2016). However, a further, linguistically and technically demanding step in such a database would be the linkage of entries on the level of individual translation equivalents among the represented languages.

In this paper, we introduce an online bilingual connective lexicon for Czech and German discourse connectives, GeCzLex, based, just like Connective-Lex, on the monolingual CzeDLex (for Czech, Mírovský et al. (2017a), updated 2019) and DiMLex (for German, the newest version Scheffler and Stede (2016)), but in addition linked on the level of individual lexicon entries, providing translation equivalents of each lexicon entry into the second language and vice versa. In its current state, the lexicon is a part of research on anaphoricity in Czech and German connectives, so it contains anaphoric connectives documented in both languages and their (not necessarily anaphoric) translation equivalents. The relevance of the translations in the lexicon is substantiated by frequencies of translation equivalents in the large collection of Czech and German parallel texts in the InterCorp project (Rosen et al., 2018). The lexicon now contains 42 Czech and 56 German anaphoric connectives and a large number of their possible translation equivalents. The chosen connectives meet either the formal or the functional criterion (or both) of anaphoricity, as defined further in Section 2.

The paper is arranged as follows: Apart from an insight into anaphoric connectives, Section 2 is also devoted to some practical decisions for the lexicon development, followed by a closer description of underlying lexicons, corpora and tools used for GeCzLex build-up in Section 3. The development procedure, which involved automatic and manual steps, is presented in Section 4. In Section 5, the structure of the lexicon is described along with an example of a lexicon entry, and we conclude with a discussion about the pros and cons of our approach and possible extensions (Sections 6 and 7).

2. Anaphoricity in Connectives

Discourse connectives serve as the most apparent anchors of discourse relations (semantico-pragmatic relations between text segments called discourse arguments), e.g., the connective *so* expresses the relation of result in Example 1 from the InterCorp corpus.

(1) The labeling is working. *It is discouraging smoking*. **<u>So</u> now Philip Morris is demanding to be compensated for lost profits.**[3]

At the same time, a coherent text is established by more aspects than discourse relations; important ties are referential

[1] http://textlink.ii.metu.edu.tr/

[2] Arabic, Bangla, Czech, Dutch, English, French, German, Italian, Portuguese

[3] In the examples in this paper, Arg1 of a discourse relation is highlighted in italics, Arg2 in bold, the connective is underlined.

relations, as demonstrated by the pronoun *this* in Example 2 from InterCorp, which anaphorically refers to the preceding sentence.

(2) The room was small, grey and humming. This was the nerve centre of the entire Guide.

In this work, we focus on a specific group of connectives where discourse relations and anaphoric relations meet, the so-called *anaphoric connectives* (ACs).
In the course of the project, it proved convenient to start working with two different definitions of anaphoric connectives, to avoid confusion: a formal one and a functional one. Language expressions and phrases complying with either of these definitions can be different but there is an intersection and both such sets of connectives are included in GeCzLex. According to the formal definition, an anaphoric connective is an expression[4] containing an anaphoric element – regarding its structure, it is usually formed from a preposition (adposition) and a referential component (e.g. *darum* in German, *proto* in Czech). Compare the connective *therefore* in Example 3 from InterCorp that expresses the relation of result and, at the same time, contains the anaphoric element *there*.

(3) *He has something to say to her*, **therefore he's come to say it**.

From the functional perspective, and in accordance with Webber et al. (2003), anaphoric connectives have, like demonstratives, the ability to relate anaphorically, not syntactically, to their left-sided argument, which also includes the possibility to relate "remotely" to non-adjacent text segments. More precisely, ACs can also accept distant text segments as their left-sided arguments, cf. Example 4 from the Prague Dependency Treebank 3.5 (Hajič et al., 2018). In this way, functionally anaphoric connectives can also contribute to higher (global) structure of the discourse.

(4) *Mám dva starobylé word-processory.* Jeden z roku 1917 a druhý z roku 1919. Vlastnoručně jsem si je natřel jasnými, tropickými barvami. **Ale na psacím stroji píšu jen některé články.**

[*I have two ancient word-processors.* One from 1917 and the other from 1919. I myself painted them in bright, tropical colors. **But I only write some articles on a typewriter.**]

On the other hand, in contrast to Webber et al. (2003), who assign this ability to adverb connectives only, we make no constraints on PoS of anaphoric connectives and base our work solely on gold discourse-annotated data. Recent corpus studies show, at least for Czech and English, that also non-adverbial connectives with no explicit anaphoric element, like *but* and its Czech counterpart *ale*, can relate to distant left-sided text segments, for exact figures compare Poláková and Mírovský (2019). That is why we included such connectives into GeCzLex, too.
In the web interface of the lexicon, for both languages, the connectives with an explicit referential element and those without it are listed in different sections and distinguished by color (blue for the former, brown for the latter). The ability to relate to non-adjacent left-sided contents was documented for all the listed connectives without a referential element (the brown ones) and for some Czech connectives with a referential element (e.g. *proto* (*therefore*), *přesto* (*yet*), *přitom* (*and/yet*) and some secondary connectives).

Individual languages differ in the ability to incorporate anaphoric elements directly into discourse connectives. While English only has a few connectives containing an anaphoric expression (e.g. *thereby*, *therefore*, *thereafter* or *whereas*), the repertoire of Czech and German is richer in this respect. We find even systemic devices for forming anaphoric connectives in German like the morphemes *da-* and *wo-* appearing, for instance, in *dagegen* and *wogegen* (*in contrast* and *whereas*). In general, German has a large number of connectives containing an anaphoric expression which is caused probably by the fact that it is a language with a strong tendency to word-formation by composition. Among German anaphoric connectives, there is a large and relatively homogeneous group of expressions consisting of an anaphoric element and a preposition like *außerdem* (*moreover*), referred to as "Pronominaladverbien" in German grammars. It is naturally assumed that connectives of this group, meeting already the formal AC definition, would also meet the functional one and relate anaphorically to their left-sided environment. This assumption is further discussed and confronted with data in Sec. 6.3.

3. Underlying Language Data and Tools

This section describes the underlying lexicons, corpora and tools that were used in order to build the current version of the bilingual GeCzLex.

CzeDLex, the Lexicon of Czech Discourse Connectives, was first published in 2017 (Mírovský et al., 2017b) and its newest version appears two years later (Synková et al., 2019). The lexicon contains 205 lemmas of Czech discourse connectives extracted from the manually annotated data of the Prague Discourse Treebank (PDiT, version 2.0, see below). CzeDLex reflects the division of connectives into primary and secondary described e.g. in Rysová and Rysová (2018), and adapted for application in lexicons in Danlos et al. (2018). All entries are annotated with basic linguistic information (a richer annotation is in progress).

The Prague Discourse Treebank is a manually annotated corpus, first published as Poláková et al. (2012), the new version appeared as Rysová et al. (2016). Subsequently, it became a part of the Prague Dependency Treebank 3.5 (Hajič et al., 2018). The corpus contains a detailed annotation of morphology, surface as well as deep syntax and the annotation of discourse phenomena (discourse connectives and relations; coreference and bridging relations). The treebank consists of almost 50 thousand sentences of contemporary Czech newspaper texts.

[4] or a multiword phrase, depending on the degree of grammaticalization

connective	sense	category
ale [*but*]	concession-arg1-as-denier	coord. conjunction
	concession-arg2-as-denier	coord. conjunction
	conjunction	coord. conjunction
	contrast	coord. conjunction
	substitution-arg2-as-subst	coord. conjunction
místo toho [*instead*]	contrast	prepositional phrase
	substitution-arg2-as-subst	prepositional phrase
totiž [*you see*]	substitution-arg2-as-subst	particle
	cause-reason	coord. conjunction
	level-of-detail-arg2-as-detail	coord. conjunction
	level-of-detail-arg1-as-detail	coord. conjunction
zatímco [*while*]	contrast	subord. conjunction
	synchronous	subord. conjunction
	conjunction	subord. conjunction

Table 1: A sample of information extracted from CzeDLex.

DiMLex, German Discourse Marker Lexicon, (Stede and Umbach (1998), Stede (2002), Scheffler and Stede (2016)) currently covers almost 300 connectives used in German. The list is synchronized with Handbuch der deutschen Konnektoren (Pasch et al., 2003) and the current version of DiMLex aims to describe all German connectives in use.

InterCorp (Rosen et al., 2018) is a multilingual parallel corpus that belongs to the family of corpora under the Czech National Corpus. InterCorp covers texts from about 40 languages; we used those in Czech and German. The German part of InterCorp (that is parallel to the Czech one) contains 6,543,622 sentences and consists, e.g., of newspaper, legal, administrative texts and fiction.

Treq (Vavřín and Rosen, 2015) is a tool for automatic searching of translation equivalents in the parallel InterCorp data. It enables to specify the desired language pair (where, currently, Czech or English serve as pivots), and subsequently to search for translation equivalents on the level of a specific word form, a lemma, a multiword unit or using regular expressions.

The Potsdam Commentary Corpus (PCC, Stede (2004)) consists of German newspaper commentaries taken from the Märkische Allgemeine Zeitung and Tagesspiegel. The corpus contains 220 commentaries (2,900 sentences, 44,000 tokens) and it is annotated with sentence syntax, coreference, discourse structure (RST), connectives, their arguments and senses and aboutness topics.

4. Method and Development Process

Providing word translations without taking semantic ambiguity into account is often insufficient. So, for example, we can get a large set of German translations for a Czech connective *ale* (*but*): *aber*, *allerdings*, *dennoch*, *doch*, *jedoch*, *sondern* and *trotzdem*, which are by no means equivalent and cannot be freely switched one for another. Our method uses available resources to overcome this deficiency and offers more precise translations with respect to possible meanings of the connectives. It uses two types of resources:

connective	sense	category
indessen	synchronous	padv, subj
	contrast	padv, subj
obwohl	concession-arg1-as-denier	subj
während	synchronous	subj, praep
	contrast	subj
wohingegen	contrast	postp
	conjunction	postp

Table 2: A sample of information extracted from DiMLex.

1. Two monolingual lexicons of discourse connectives (DiMLex for German connectives, CzeDLex for Czech connectives) that for each connective provide possible senses (discourse types) the connective can express (see 4.2.1).

2. Two bilingual translation tables that provide translation candidates of connectives for both directions (Czech–German, German–Czech), see 4.2.3.

Following, for instance, the Czech–German translation direction, the method proceeds with the following steps:

1. For each Czech connective, senses (discourse types) expressible by the connective are retrieved from CzeDLex and listed as possible readings of the connective.

2. German candidate translations of the connective are retrieved from the Czech–German translation table; then, for each sense (discourse type) from step (1), only those candidates that – according to DiMLex – can express the given sense are selected as valid German translations of the Czech connective in the given sense.

In practice, we have applied step (1) of the procedure in both translation directions (Czech–German and German–Czech) to anaphoric connectives only, while in step (2), translation candidates can be connectives of any kind (not only anaphoric).

4.1. Combining the Resources

Let us exemplify the procedure on the Czech connective *zatímco* (*while*). Relevant samples of the key resources needed in the Czech–German translation direction are depicted in Tables 1 and 2, representing information from CzeDLex and DiMLex, respectively, and Table 3, representing the connective translation candidates table from Czech to German. (Section 4.2 gives details about the preparation of these resources.)

1. Table 1 (i.e. CzeDLex) offers three possible senses for the connective *zatímco*: contrast, synchronous and conjunction.

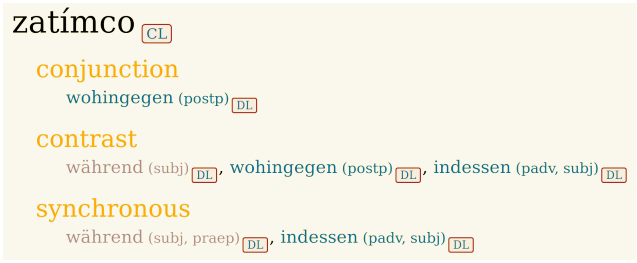

Figure 1: The GeCzLex entry for the connective *zatímco* (*while*). The senses are ordered alphabetically (i.e. not according to corpus counts).

connective	translation candidates
ale [*but*]	*aber*, *allerdings*, *dennoch*, *doch*, *jedoch*, *sondern*, *trotzdem*
nadto [*apart from that*]	*überdies*, *außerdem*, *zudem*, *darüber hinaus*, *obendrein*, *zumal*
také [*also*]	*auch*, *ebenfalls*, *außerdem*, *ferner*, *zudem*, *ebenso*, *darüber hinaus*
zatímco [*while*]	*während*, *obwohl*, *wohingegen*, *indes*, *indessen*

Table 3: A sample from the Czech–German translation candidates table.

2. The Czech–German translation table (Table 3) offers five German translation candidates: *während*, *obwohl*, *wohingegen*, *indes*, *indessen*.

3a. List of senses expressible by connective *während* taken from Table 2 (i.e. DiMLex) shows that connective *während* is a suitable translation for senses contrast (with the morpho-syntactic category subj) and synchronous (with categories subj and praep) but not for the sense conjunction, as – according to DiMLex – *während* cannot express this sense.

3b. Similarly, Table 2 helps assign translation candidates *indessen* and *wohingegen* to the appropriate senses.

3c. Translation candidate *obwohl* is not assigned to any sense (and therefore dismissed), as – according to DiMLex (Table 2) – it can only express the sense concession-arg1-as-denier.

3d. Translation candidate *indes* is also dismissed, as it is not in DiMLex.

Figure 1 shows the resulting GeCzLex entry for the Czech connective *zatímco*.

4.2. Preparation of the Resources

4.2.1. Extraction of Senses

From both monolingual lexicons of discourse connectives, for each connective, the list of possible senses is extracted, along with the morpho-syntactic category(ies) associated with the connective and the given sense. We use DiMLex in its native XML format; CzeDlex has been first transformed to a DiMLex-like format and its Prague taxonomy of discourse relations has been translated to the PDTB 3 taxonomy (cf. details on mapping the taxonomies in 4.2.2). Tables 1 and 2 show a few examples of such extracted information from CzeDLex and DiMLex, respectively. The morpho-syntactic category (PoS) is not used yet in mapping the translations but is later printed as an additional information next to the translated connectives.

4.2.2. Sense Mapping

The original German DiMLex from 1998 had no semantic information included. In 2016, the lexicon was substantially widened as well as enriched with semantic relations (Scheffler and Stede, 2016) that were adopted from the freshly established tagset of Penn Discourse Treebank version 3 (PDTB 3 tagset, Webber et al. (2016)). This taxonomy, a result of feedback by many discourse researchers working with earlier PDTB tagsets or similar taxonomies, also became the common tagset for the multilingual Connective-Lex, as mentioned earlier. In the case of the Czech lexicon CzeDLex, extracted from Czech annotations that used a modified PDTB 2 semantic taxonomy, this meant to find a mapping between the Prague relations and the recent PDTB 3 relations. Within the four major semantic classes (Temporal, Contingency, Comparison, Expansion), which are the same for both tagsets, there were no issues in mappings. Also, some new relations in the PDTB (e.g., level-of-detail vs. specification/generalization) caused no trouble. At a more fine-grained level, there was some loss of information: for instance, the Czech semantic type "gradation"[5] has no direct counterpart in PDTB 3 and so it was mapped onto the broader PDTB 3 "conjunction". In the opposite direction, Prague labels do not include "manner", which was merged with "specification", or in other words "level-of-detail:Arg2 as detail", or "negative condition" – this label was introduced in newly annotated data only recently, and thus it is temporarily merged with Prague "condition". The resulting sense mapping table became the underlying material for sense extraction described above in Sec. 4.2.1.

4.2.3. Translation Candidates Tables

As stated earlier, German exhibits a stronger tendency to word-formation by composition than Czech, which is also projected in the form of anaphoric connectives. German contains more grammaticalized (single-word) anaphoric connectives than Czech. Taking into account semantic equivalents in Czech and German, we observe that many anaphoric connectives, which appear as single words in German are multiword phrases in Czech. Therefore, in the Czech part of the lexicon, we also cover multiword phrases corresponding to the structure "preposition + anaphoric element". These phrases include e.g. *kvůli tomu* (*because of this*) or *místo toho* (*instead*) which mostly have single-word counterparts in German: *deswegen* and *stattdessen*, respectively). In this way, we selected three groups of formally anaphoric connectives: grammaticalized connectives in German (altogether 54 connectives), grammaticalized connectives in Czech (17) and non-grammaticalized con-

[5] the typical *not only... but also* or *moreover* meaning

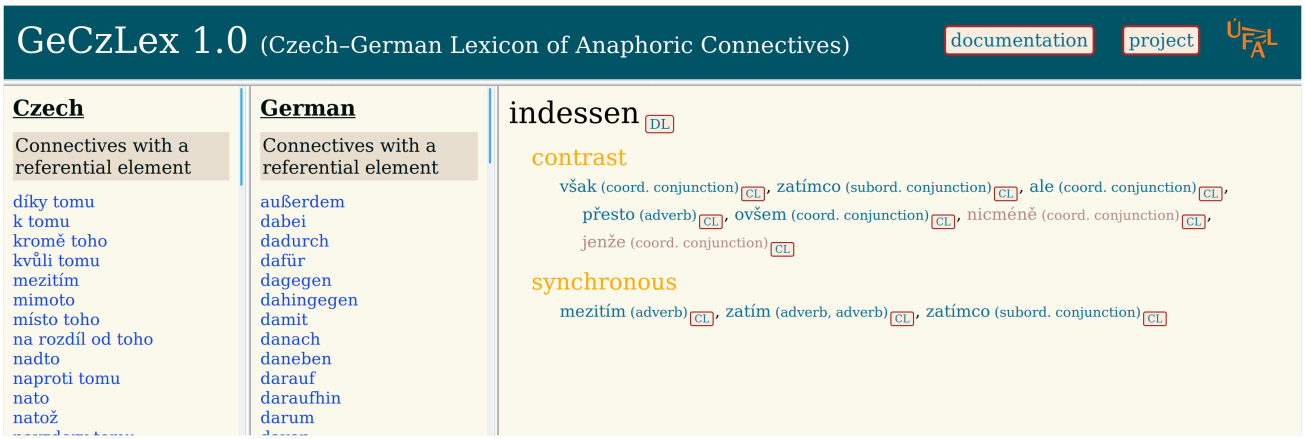

Figure 2: A screenshot of GeCzLex with the entry for German connective *indessen* selected.

nectives (multi-word phrases) in Czech (11).
An earlier PDiT 2.0 querying returned further 14 Czech connectives with no explicit anaphoric element but with remote left-sided argument; studying the (rather small) PCC data revealed two such connectives for German, cf. 6.3.
Then, using the Treq tool on the parallel texts in InterCorp, we manually created translation candidate tables with lists of most common equivalents for each AC in both languages. Treq also produces their percentage; that is why we could include only stable and justified translations and omit possible alignment errors. The resulting lists were sorted according to translation frequency and manually filtered for non-connective readings. Table 3 gives examples of translations candidates in the Czech–German direction.

5. Lexicon Content and Structure

In the web HTML interface (see Figure 2), the two columns on the left represent the lists of Czech and German connectives included in the lexicon at present, respectively. By clicking on an item, the lexicon entry for the given connective (*indessen* in Figure 2) opens in the main frame. The current GeCzLex lexicon entry contains:

- The entry head – **the lemma of the connective** (e.g., *indessen*). For Czech secondary connectives, unlike in CzeDLex, the lemma is not just the core word of the phrase, but it is the whole (prepositional) phrase so that the referential component is visible at first sight.
- **The URL link to the full entry** of the given connective in the underlying resource, that is CL - CzeDLex for Czech connectives and DL - DiMLex for German connectives. The URL link for German connectives actually leads to the connective's representation in Connective-Lex.[6] Thanks to the linkage with these underlying online inventories, the user can easily access exhausting information about the connective together with examples (original corpus examples in case of Czech connectives). Such a link is then provided also for every translation of a given connective.
- For each entry lemma, a list of assigned **semantic relations (senses)** from the PDTB 3 tagset is displayed (for connective *indessen* in Figure 2, contrast and synchronous). At present, the ordering of the semantic relations in the lexicon entry is alphabetical, not sorted according to the corpus frequencies. For each sense, a list of translations with the same sense label in the second language is given (in Figure 2, e.g., for the sense synchronous: *mezitím*, *zatím*, *zatímco*). The translations are sorted in the descending order according to the frequency in the source parallel corpus.
- **Intra-lexicon links**: if a translation of a given connective is also an anaphoric connective, clicking on its lemma again opens up its GeCzLex entry. If it is not, it is displayed in a different color and does not contain a hypertext link.
- For each translation, **syntactic categories**, i. e. part of speech (for Czech primary connectives) or syntactic structure (for German connectives) are extracted from the underlying lexicons, and, in case of secondary connectives – multiword structures in Czech, their syntactic structure is added manually.[7] This should help the user to notice possible syntactic constraints in using a given translation immediately.

6. Discussion

The current version of GeCzLex is the first attempt of such an electronic inventory, which brings along the need to resolve many newly emerged questions, both technical and linguistic.
First, we acknowledge that in the current state of development, the choice of an appropriate translation equivalent for

[6] As there is no way to make a link to a specific Connective-Lex entry (a single connective), we employ the possibility to encode a search query on top of the whole Connective-Lex in the URL link. We search for the required connective as a string among connectives in the German part of Connective-Lex. This way, usually just the required single entry is retrieved; occasionally, superfluous entries are listed as well – in case the connective is a substring of other connectives too (e.g., entries for *darauf* and *daraufhin* are retrieved from Connective-Lex when asked for *darauf*).

[7] as CzeDLex offers PoS for the core word of the phrase only

a given context from a GeCzLex entry can be further influenced by the morphosyntactic and word order constraints of the connectives, register and style. However, at least syntactic categories for all connectives are provided in GeCzLex, and further context-relevant information, including use examples, can be found through the links to the underlying lexicons. The Czech examples in the Czech lexicon are authentic corpus examples and they are manually translated into English.

Next, in the rest of this Section, we address some general issues encountered during the lexicon compilation in more detail. We believe our considerations may be helpful for other researchers aiming at any such lexicon build-up.

6.1. Translation Asymmetry

In order to capture as many translation possibilities of ACs as possible, we initially considered to automatically add translation equivalents from the opposite translation direction in cases where the backtranslation did not include the original expression. In other words, we wanted to make the lexicon symmetric. However, as all translation possibilities of a given connective in GeCzLex had to be documented in parallel corpora with a relevant number/percentage of occurrences, there must have been a systematic reason why the lexicon entries were asymmetric in translation. In particular, this holds for translation equivalents with a different degree of semantic granularity. A nice example is the Czech connective *díky tomu* (*thanks to that*):[8] it is quite commonly translated as the semantically more general *deswegen* or *dadurch* (*therefore*), but, in the opposite direction, these German connectives are rarely translated as *díky tomu*. As these disproportions are difficult to identify and support by numbers, automatic global addition of the opposite translations would lead to proportional mismatches of possible translation equivalents in the lexicon compared to InterCorp data or even to non-documented translations.

Abandoning this method of lexicon enrichment and also the limited range of the underlying monolingual connective lexicons resulted in the fact that some semantic types in some entries render no translations at all. This is, in our opinion, not a deficiency of the resource but, on the contrary, valuable linguistic feedback on connective repertoires and translation in the two languages, their semantic categorization and coverage of the two underlying lexicons. All these types of feedback can help further enhance the underlying data/lexicons.

6.2. Cataphoric Connectives

Some connectives with a referential element also have the ability to express a cataphoric relation simply by introducing a subordinate clause, cf. the Czech connective pair *vzhledem k tomu* (*therefore*, *as a result*, lit. "*with respect to this*") and *vzhledem k tomu, že* (*because*, lit. "*with respect to the fact that*"). The distinction between anaphoric vs. cataphoric use of these connectives is illustrated in Examples 5 and 6 from PDiT, the former variant shows an anaphoric relation (semantically result), the latter signals a cataphoric relation (semantically reason).

(5) Jenomže v těchto zemích sami nevědí, co s nezaměstnanými, odpověděl Otto Brabec z agentury Servus na otázku, jak se daří zprostředkovávat práci v zahraničí. Vzhledem k tomu (anaphoric use) dominuje v činnosti agentury zajišťování studentských pracovních pobytů především v Německu.

[But in these countries, they themselves do not know what to do with the unemployed, Otto Brabec of the Servus agency replied to the question of how they manage to mediate work abroad. As a result, in the activities of the agency, mediation of student's work dominates mostly in Germany.]

(6) Je nejvyšší čas hledět dopředu i vzhledem k tomu, že (cataphoric use) většina našich občanů se narodila již po válce.

[It is high time to look ahead also because most of our citizens were born after the war.]

The referential element of these connectives (most commonly a demonstrative pronoun *to* (*this/that*)) refers either backwards or forwards into the text. The anaphoricity or cataphoricity of the referential element may also affect the semantics of the discourse relation, more precisely the direction of asymmetric relations (cf. reason vs. result in Examples 5 and 6).

The ability to refer cataphorically is given especially to secondary connectives – multiword phrases like *kvůli tomu, že* (*because*, lit. "*due to the fact that*"), *kromě toho, že* (*besides*, lit. "*beside the fact that*") etc. However, cataphoric links are observable also in primary connectives, cf. the connective *předtím* (*before that*) and a cataphoric use of *předtím, co* lit. "*before that that*" expressing the relation of precedence. Possible ways of treatment of cataphoric connectives in GeCzLex are still under debate.

6.3. Non-Adjacency and German Connectives

The original research question was driven by the hypothesis that connectives with an explicit referential element, which the German "Pronominaladverbien" largely are, are more likely to introduce relations with a non-adjacent left (external) argument. However, Stede and Grishina (2016) in their work on German anaphoric connectives report that the absence of an explicitly anaphoric morpheme in the connective does not exclude its anaphoric behavior. We can support this claim by having detected German connectives in the discourse-annotated data of Potsdam Commentary Corpus that take a non-adjacent external argument, cf. Example 7. They are *dann* (*then*) and *allerdings* (*however*). These two connectives do not agree with the formal AC definition. The other way round, there were no instances detected of a German pronominal adverb with a distant external argument. Nevertheless, it must be emphasized, the corpus data of the PCC is probably not large enough to document the possible anaphoric behavior of these expressions.

(7) Um die deutschen Legehennen ist heftiger politischer Streit entbrannt. *Bundesagrarministerin Renate Künast will das Halten der Tiere in engen Legebatterien bereits vom Jahr 2006 an verbieten.* In den

[8] apart from the most fitting German *dank*, which does not form an AC in German and cannot be used in all syntactic settings

EU-Nachbarländern soll das erst fünf Jahre später gelten. Eier-Produzenten aus der ganzen Republik machen gegen Künasts Pläne mobil. Die Betriebe im Osten fürchten, dass die hohen Investitionen, die sie in moderne Legebatterien nach europäischem Standard gesteckt haben, umsonst gewesen sind. **Im Westen herrscht die Sorge vor, ausländische Konkurrenten könnten dann mit Billig-Eiern aus Legebatterien den deutschen Markt überschwemmen.**

7. Conclusion

We have introduced a new bilingual resource describing translations between Czech and German discourse connectives, the GeCzLex. The lexicon was built via interlinking existing monolingual lexicons and validating the translation counterparts on a large collection of parallel data. Such a resource aims to keep track of the correct, semantics-sensitive translation of discourse connectives, usable both for human users and translation systems. The lexicon underwent a great deal of manual enhancement, that is still in progress. The development version of GeCzLex is available on-line as HTML web pages.[9] GeCzLex was officially released at the Lindat/Clarin repository[10] under the Creative Commons License (Rysová et al., 2019). At present, the lexicon contains primarily anaphoric Czech and German connectives, as our goal was to closer investigate their referential potential connected with the ability to relate distant (non-adjacent) text segments and possibly play a role in global discourse coherence. The linking procedure, however, is easily expandable to all connectives that are covered by CzeDLex and DiMLex. Tables with translation candidates might be obtained automatically from word-alignments of Czech–German parallel resources; although such tables would contain much noise, most errors would likely be filtered out by cross-referencing with CzeDLex and DiMLex. In future, finer information from both lexicons can be added to GeCzLex and the entries can be further accompanied with authentic parallel corpus examples.

8. Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (projects GA17-06123S and GA20-09853S). This work has been supported by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

9. References

Danlos, L., Rysová, K., Rysová, M., and Stede, M. (2018). Primary and Secondary Discourse Connectives: Definitions and Lexicons. *Dialogue and Discourse*, 9(1):50–78.

Das, D., Scheffler, T., Bourgonje, P., and Stede, M. (2018). Constructing a lexicon of english discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365.

Feltracco, A., Jezek, E., Magnini, B., and Stede, M. (2016). Lico: A Lexicon of Italian Connectives. *CLiC it*, page 141.

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Hladká, B. V., Zeman, D., Zikánová, Š., and Žabokrtský, Z. (2018). *Prague Dependency Treebank 3.5*. Univerzita Karlova, MFF, ÚFAL, Prague, Czech Republic.

Mendes, A. and Lejeune, P. (2016). LDM-PT-A Portuguese Lexicon of Discourse Markers. In *Conference Handbook of TextLink–Structuring Discourse in Multilingual Europe Second Action Conference*, pages 89–92. Debrecen University Press.

Mírovský, J., Synková, P., Rysová, M., and Poláková, L. (2017a). CzeDLex–A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, 109(1):61–91.

Mírovský, J., Synková, P., Rysová, M., and Poláková, L. (2017b). *CzeDLex 0.5*. Charles University, Prague, Czech Republic.

Poláková, L. and Mírovský, J. (2019). Anaphoric Connectives and Long-Distance Discourse Relations in Czech. *Computación y Sistemas*, 23(3).

Poláková, L., Jínová, P., Zikánová, Š., Hajičová, E., Mírovský, J., Nedoluzhko, A., Rysová, M., Pavlíková, V., Zdeňková, J., Pergler, J., and Ocelák, R. (2012). *Prague Discourse Treebank 1.0*. ÚFAL MFF UK, Prague, Czech Republic.

Rosen, A., Vavřín, M., and Zasina, A. J. (2018). *Korpus InterCorp*. Ústav Českého národního korpusu, Praha.

Rysová, M. and Rysová, K. (2018). Primary and Secondary Discourse Connectives: Constraints and Preferences. *Journal of Pragmatics*, 130:16–32.

Rysová, M., Synková, P., Mírovský, J., Hajičová, E., Nedoluzhko, A., Ocelák, R., Pergler, J., Poláková, L., Pavlíková, V., Zdeňková, J., and Zikánová, Š. (2016). *Prague Discourse Treebank 2.0*. ÚFAL MFF UK, Prague, Czech Republic.

Rysová, K., Poláková, L., Rysová, M., and Mírovský, J. (2019). Lexicon of Czech and German Anaphoric Connectives. In print.

Scheffler, T. and Stede, M. (2016). Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1008–1013, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Stede, M. and Grishina, Y. (2016). Anaphoricity in connectives: A case study on German. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 41–46, San Diego, California, June. Association for Computational Linguistics.

Stede, M. and Umbach, C. (1998). DiMLex: A Lexicon of

[9] http://ufal.mff.cuni.cz/geczlex/

[10] http://hdl.handle.net/11234/1-3075

Discourse Markers for Text Generation and Understanding. In *Proceedings of the 17th International Conference on Computational Linguistics (Coling 1998)*, pages 1238–1242. Association for Computational Linguistics.

Stede, M., Scheffler, T., and Mendes, A. (2019). 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). Available from: http://connective-lex.info/.

Stede, M. (2002). DiMLex: A Lexical Approach to Discourse Markers.

Stede, M. (2004). The potsdam commentary corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102.

Synková, P., Poláková, L., Mírovský, J., and Rysová, M. (2019). *CzeDLex 0.6*. Charles University, Prague, Czech Republic. In print.

Vavřín, M. and Rosen, A. (2015). *Treq*. Ústav Českého národního korpusu, Praha.

Webber, B., Stone, M., Joshi, A., and Knott, A. (2003). Anaphora and Discourse Structure. *Computational linguistics*, 29(4):545–587.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2016). A Discourse-Annotated Corpus of Conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31, Berlin, Germany, August. Association for Computational Linguistics.