Chapter 2

Discourse connectives: From historical origin to present-day development

Magdaléna Rysová
Charles University, Faculty of Mathematics and Physics

The paper focuses on the description and delimitation of discourse connectives, i.e. linguistic expressions significantly contributing to text coherence and generally helping the reader to better understand semantic relations within a text. The paper discusses the historical origin of discourse connectives viewed from the perspective of present-day linguistics. Its aim is to define present-day discourse connectives according to their historical origin through which we see what is happening in discourse in contemporary language. The paper analyzes the historical origin of the most frequent connectives in Czech, English and German (which could be useful for more accurate translations of connectives in these languages) and point out that they underwent a similar process to gain a status of present-day discourse connectives. The paper argues that this historical origin or process of rising discourse connectives might be language universal. Finally, the paper demonstrates how these observations may be helpful for annotations of discourse in large corpora.

1 Introduction and motivation

Currently, linguistic research focuses often on creating and analyzing big language data. One of the frequently discussed topics of corpus linguistics is the annotation of discourse carried out especially through detection of discourse connectives. However, discourse connectives are not an easily definable group of expressions. Linguistic means signaling discourse relations may be conjunctions like *but*, *or* etc., prepositional phrases like *for this reason*, fixed collocations like

Magdaléna Rysová. 2017. Discourse connectives: From historical origin to present-day development. In Katrin Menzel, Ekaterina Lapshinova-Koltunski & Kerstin Kunz (eds.), *New perspectives on cohesion and coherence*, 11–35. Berlin: Language Science Press. DOI:10.5281/zenodo.814460

as seen, *simply speaking* etc., i.e. expressions with a different degree of lexicalization, syntactic integration or grammaticalization. Therefore, the paper concentrates on formulating clear boundaries of discourse connectives based on a deep linguistic research.

The paper analyzes the historical origin of the most frequent present-day connectives (mainly in Czech in comparison to other languages like English and German) to observe their tendencies or typical behaviour from a diachronic point of view, which may help us in annotation of connectives in large corpora (mainly in answering the question where to state the boundaries between connectives and non-connectives that could significantly facilitate the decision which expressions to capture in the annotation and which not). In other words, the paper tries to answer what we can learn from discourse connective formation and historical development and what this may tell us about present-day structuring of discourse.

The need for a clearly defined category of discourse connectives in Czech arose mainly during the annotation of discourse relations in the Prague Discourse Treebank (PDiT) pointing out several problematic issues. One of the most crucial was where and according to which general criteria to state the boundaries between connectives and non-connectives as well as between explicitness and implicitness of discourse relations. An explicit discourse relation is usually defined as a relation between two segments of text that is signaled by a particular language expression (discourse connective), typically by conjunctions like *a* 'and', *ale* 'but', *nebo* 'or' etc. However, during the annotations, we had to deal with examples of clear discourse relations expressed by explicit language means that, however, significantly differed from those typical examples of connectives. Such means included multiword phrases often having the function of sentence elements (like *kvůli tomu* 'due to this', *z tohoto důvodu* 'for this reason', *hlavní podmínkou bylo* 'the main condition was', *stejným dechem* 'in the same breath' etc.). Therefore, it was necessary to answer the question whether such expressions may be also considered discourse connectives and therefore included into the annotation of the PDiT or not.

It appeared that it is very helpful to look for the answer in the historical origin of the present-day typical connectives, i.e. expressions that would be without doubt classified as discourse connectives by most of the authors (like the mentioned conjunctions *a* 'and', *ale* 'but', *nebo* 'or' and many others). The results of such research (combined with the analysis of the present-day corpus data) are presented in this paper.

2 Theoretical discussions on discourse connectives

Discourse connectives are in various linguistic approaches defined very differently, which is mainly due to their complexity and hardly definable boundaries. There are several definitions highlighting different language aspects of discourse connectives – concerning their part-of-speech membership, lexical stability, phonological behaviour, position in the sentence etc. Most of the authors agree on defining the prototypical examples of connectives, i.e. expressions like *but*, *while*, *when*, *because* etc. and differ especially in multiword collocations like *for this reason*, *generally speaking* etc. The prototypical connectives are usually defined as monomorphemic, prosodically independent, phonologically short or reduced words (see Zwicky 1985; Urgelles-Coll 2010) that are syntactically separated from the rest of the sentence (see Schiffrin 1987; Zwicky 1985), not integrated into the clause structure (see Urgelles-Coll 2010) and that usually occupy the first position in the sentence (see Schiffrin 1987; Zwicky 1985; Schourup 1999; Fischer 2006).

Considering part-of-speech membership, some authors classify connectives as conjunctions (both subordinating and coordinating), prepositional phrases and adverbs (see Prasad et al. 2008; Prasad, Joshi & Webber 2010), others also as particles and nominal phrases (see Hansen 1998; Aijmer 2002), others include also some types of idioms (like *all things considered*, see Fraser 1999).

However, some of the mentioned syntactic classes (like prepositional phrases or nominal phrases) do not correspond to the definitions of discourse connectives stated above, i.e., for example, that connectives are usually short, not integrated into clause structure etc. Some of the authors define discourse connectives in a narrow sense (see e.g. Shloush 1998; Hakulinen 1998; Maschler 2000 who limit connectives only to synsemantic, i.e. grammatical words), some in a broader sense (e.g. according to Schiffrin 1987, discourse relations may be realized even through paralinguistic features and non-verbal gestures).

This paper contributes to these discussions on discourse connectives and looks at them from the diachronic point of view. It argues that the historical development of discourse connectives may point out many things about general tendencies in present-day structuring of discourse.

3 Methods and material

The analysis of discourse connectives in Czech is carried out on the data of the Prague Discourse Treebank 2.0 (PDiT; Rysová et al. 2016), i.e. on almost 50 thousand annotated sentences from Czech newspaper texts. The PDiT is a multilayer

annotated corpus containing annotation on three levels at once: the morphological level, the surface syntactic level (called analytical) and the deep syntactico-semantic level (called tectogrammatic). At the same time, the PDiT texts are enriched by the annotation of sentence information structure[1] and various discourse phenomena like coreference and anaphora and especially by the annotation of explicit discourse relations (i.e. relations expressed by concrete language means, not implicitly).

The annotation of discourse relations in the PDiT (based on a detection of discourse connectives within a text) does not use any pre-defined list of discourse connectives (as some similar projects – see, e.g., Prasad et al. 2008). The human annotators themselves were asked to recognize discourse connectives in authentic texts. Therefore, a need for an accurate delimitation of discourse connectives arose, especially for stating the boundaries between connectives and non-connectives.

The most problematic issue appeared to be the multiword phrases like *to znamená* 'this means', *výsledkem bylo* 'the result was', *v důsledku toho* 'in consequence', *podmínkou je* 'the condition is' etc. These phrases clearly signal discourse relations within a text (e.g. *podmínkou je* 'the condition is' expresses a relation of condition), but they significantly differ (in lexico-syntactic as well as semantic aspect – see Rysová 2012) from the "prototypical", lexically frozen connectives like *ale* 'but' or *a* 'and' (these phrases may be inflected, appear in several variants[2] in the text etc. – see e.g. *za této podmínky* 'under this condition' vs. *za těchto podmínek* 'under these conditions', *závěrem je* 'the conclusion is' vs. *závěrem bylo* 'the conclusion was').

At the same time, some typical Czech connectives like *proto* 'therefore', *přesto* 'in spite of this' etc. were historically also multiword – they are frozen prepositional phrases (raised from the combination of preposition *pro* 'for' with the pronoun *to* 'this' and the preposition *přes* 'in spite of' with the pronoun *to* 'this'), so the main difference between them and present-day phrases like *kvůli tomu* 'due to this' is that they are now used as one-word expressions. This idea raises many questions – e.g. is the frozen lexical form (that appears in most of the typical present-day connectives in Czech) a sufficient argument to exclude the multiword phrases from discourse connectives and their annotation in the corpus? Would not the annotation without them be incomplete?

This led us to the idea to examine the historical origin of other 'prototypical' discourse connectives in Czech, which could tell us something about the men-

[1] To sentence information structure in Czech see, e.g., Hajičová, Partee & Sgall (2013) or Rysová (2014a).

[2] See also a study on reformulation markers by Cuenca (2003).

tioned multiword phrases in general and could suggest their uniform annotation in the corpus. In this respect, the paper concentrates on where to put the boundaries of discourse connectives so that the annotations of large corpus data are not incomplete and at the same time follow an adequate theoretical background.

4 Results and evaluation

4.1 Historical origin of the most frequent connectives in Czech

In these subsections, the paper presents the results of the analysis of discourse connectives with emphasis on their historical origin and development towards their present-day position in language. In this way, the paper introduces a comparative study of Czech, English and partly German.

Table 1: Most frequent Czech connectives in the PDiT

Czech connectives	Tokens in the PDiT
a 'and'	5,765
však 'however'	1,521
ale 'but'	1,267
když 'when'	574
protože 'because'	525
totiž 'that is'	460
pokud 'if'	403
proto 'therefore'	380
tedy 'so'	307
aby 'so that'	305

For the analysis, the ten most frequent discourse connectives in Czech (presented in Table 1) have been selected and their historical origin have been analyzed – see Table 2[3].

Table 2 demonstrates that none of the selected connectives was a connective from its origin. All of them arose from other parts of speech than conjunctions or structuring particles or from a combination of several words. At a certain

[2] The Czech connective *totiž* does not have an exact English counterpart; a similar meaning is carried by the German *nämlich*.

[3] The etymology of Czech connectives is adopted from the Czech etymological dictionaries and papers (see Holub & Kopečný (1952); Rejzek (2001); Bauer (1962); Bauer (1963)).

moment, this word or words began to be used in a connecting function, which started the process of their grammaticalization (cf. related works by Claridge & Arnovick 2010; Degand & Vandenbergen 2011; Claridge 2013 or Degand & Evers-Vermeul 2015).

This process began for the individual connectives in different periods (one of the oldest seems to be the rise of *a* 'and' in Czech as similarly *and* in English and *und* 'and' in German – see below). Sometimes the grammaticalization is not fully completed, which causes the discrepancies within some parts of speech (in Czech mainly within adverbs, particles and conjunctions). The unfinished grammaticalization is seen, e.g., on connectives that are still written as two words (like Czech *a tak* 'and so', *i když* 'even though' etc.) in contrast to already one-word connectives containing historically the same component *a* 'and' – *ale* 'but', *ač* 'although', *aby* 'so that'.

Table 2 shows that Czech present-day most frequent connectives originally arose from other parts of speech than, e.g., conjunctions, i.e. they are not connectives from their origin, but they gained a status of connectives during the historical development. Some of the Czech connectives arose from interjections (e.g. *a* 'and'), adverbs (e.g. *však* 'however') or adjectives (e.g. *také* 'too'). Most of them are originally compounds of two components (mainly interjections, particles, adverbs or prepositions). Some of the combinations even repeat – see combinations of preposition and pronoun (*pro-to* 'therefore', *při-tom* 'yet', *o-všem* 'nevertheless'), pronoun and particle (*te-dy* 'so', *co-ž* 'which') or preposition and adverb (*po-kud* 'if', *na-víc* 'moreover').

Some of the connectives are even combinations of three components – like preposition, pronoun and particle (*pro-to-že* 'because') or preposition and two pronouns (*za-tím-co* 'while'). Therefore, it is evident that the most frequent Czech connectives were (before they became one-word expressions) very similar to the present-day multiword phrases like *kvůli tomu* 'due to this' or *z tohoto důvodu* 'for this reason'. The origin of some of them is rather transparent even today (e.g. most native speakers are probably able to recognize that the connective *proto* 'therefore' is a compound of preposition *pro* 'for' and a pronoun *to* 'this') while some of them have (synchronically) lost motivation (see mainly the oldest connectives like *ale* 'but', *nebo* 'or' etc.). This fact is depending on the degree of their grammaticalization – the more grammaticalized the connective is, the less bonds remain to its historical origin. In this respect, discourse connectives are not a closed class of expressions, but rather a scale representing the process of connective grammaticalization.

Table 2: Historical origin of most frequent discourse connectives in Czech

Czech present-day connectives	Historical origin
a 'and'	from a deictic interjection meaning *hle* 'behold'
však 'however'	adverbial origin meaning 'always'
ale 'but'	combination of *a* 'and' (with interjectional origin) and particle *-le* (with the adverbial meaning *jen* 'only')
když 'when'	combination of adverb *kdy* 'when' and particle *-ž (že)* (today's conjunction 'that')
protože 'because'	combination of three components: preposition *pro* 'for', pronoun *to* 'this' and particle *-ž (že)* (today's conjunction 'that')
totiž 'that is'	unclear origin: either combination of three components: pronoun *to* 'this', particle *-ť (ti)* and particle *-ž (že)* (today's conjunction 'that') or grammaticalized verbal phrase *točúš/točíš* [lit. (you) it know] coming from the composition of a demonstrative pronoun *to* 'this' and a verb *čúti/číti*
pokud 'if'	combination of preposition *po* 'after' and adverb *kudy* 'from where'
proto 'therefore'	combination of preposition *pro* 'for' and pronoun *to* 'this'
tedy 'so'	combination of pronoun *to* 'this' and particle *-dy (-da)*
aby 'so that'	combination of *a* 'and' and verbal component *bych* (derived from the verb *být* 'be')

The given expressions in certain combinations and in certain forms begun to be used as connectives and they underwent the process of grammaticalization (in different time period) – thus, the individual present-day connectives lay in different parts of the scale according to the degree of their grammaticalization.

4.2 Historical origin of the most frequent connectives across languages

We have compared the results of analysis of Czech connectives with their counterparts in English[4] to see whether the connectives in another language exhibit similar behaviour – see Table 3.

Table 3[5] demonstrates that the origin of given English connectives is very comparable to their Czech counterparts. Also English connectives are not connectives from their origin. They arose also from other parts of speech (mainly from combinations of pronouns, prepositions and adverbs) or other multiword phrases. Many of them (not only presented in Table 3) have a pronominal origin (like *when*, *if*, *so*, *then*, *which*), many come from the whole phrases that may have two or more components – see the combination of an adverb and pronoun (*how-ever*) or adverb and preposition (*there-fore*).

Similar connective formation may be seen also in German.[6] For example, the connective *dass* 'so that, that' arose from a demonstrative pronoun *das* 'this', *jedoch* 'however' from the combination of two words: *je* 'sometimes' and conjunction *doch* 'however'.

The connective *nämlich* 'that is' (a counterpart to Czech *totiž*) is historically an unstressed variant of an adverb *name(nt)lich* 'namely' derived from the noun *Name* 'name'; the original meaning of *nämlich* is 'the same' but it shifted to present-day more often adverbial meaning of 'it means, more specifically'. The semantic shift is seen also in other German present-day connectives like *weil* 'because' (today, with a causal meaning, but originally expressing a temporal relation – cf. the German noun *Weile* 'moment' or English temporal conjunction *while*), *aber* 'but' (originally expressing multiple repetition like 'once again, again'), *wenn* 'when, if' (originally an unstressed variant of *wann* 'when' with

[4] Apart from the Czech connective *totiž* that does not have an appropriate counterpart in English (but it roughly corresponds to German connective *nämlich*).

[5] The etymology of English connectives is adopted from the English etymological dictionary – Harper (2001). The aim of this paper is not to discuss the etymology of English connectives in general (which is in detail in Lenker & Meurman-Solin (2007)), but to compare the origin of some of them with their Czech counterparts.

[6] The etymology of German connectives is adopted from Klein & Geyken (2010).

Table 3: Historical origin of selected discourse connectives in English

English present-day connectives	Historical origin
and	Old English *and*, *ond*, originally meaning 'thereupon, next' from Proto-Germanic *unda
however	combination of *how* and *ever* (late 14th century)
but	combination of West Germanic *be- 'by' and *utana 'out, outside, from without'; not used as conjunction in Old English
when	from pronominal stem *hwa-, from PIE interrogative base *kwo
because	combination of preposition *bi* and noun *cause*: *bi cause* 'by cause', often followed by a subordinate clause introduced by *that* or *why*; one word from around 1400
if	coming from Proto Indo-European pronominal stem *i-
therefore	combination of *there* and a preposition *fore* (an Old English and Middle English collateral form of the preposition *for*) meaning 'in consequence of that'
so	from Proto Indo-European reflexive pronominal stem *swo-, pronoun of the third person and reflexive
so that	unmerged conjunction of two components

temporal meaning; today, it expresses both temporal as well as conditional relations) etc.

A large group of present-day connectives arose from combination of prepositions and a deictic component *da* – see the so called anaphoric connectives like *dafür* lit. 'for this/that', *davor* 'previously', *danach* 'then', *darum* 'therefore' etc.

We see that the general principle of discourse connectives development was very similar in Czech, English as well as German. Therefore, it may be supposed that formation of discourse connectives is not language specific but language universal.

5 Formation of discourse connectives

5.1 General tendencies

In this part, the paper summarizes the most frequent formations for present-day discourse connectives (with more examples as well as from other languages) to demonstrate that there are some productive connective formations across the languages' development.

Firstly, the paper summarizes the general tendencies for connective formation in Czech. During the analysis above, we could observe that many of the Czech connectives follow similar principles and in some cases, they are formed even by the same components – see the following five points.

1. One of the most productive components (forming the final part of many Czech connectives) is the particle *-ž(e)*[7] occurring in the grammaticalized one-word connectives as well as in unmerged multiword phrases – see one-word examples like *což* 'which', *protože* 'because', *když* 'when', *též* 'too', *než* 'than', *nýbrž* 'but', *tudíž* 'thus', *až* 'until', *poněvadž* 'because', *jelikož* 'because', *jestliže* 'if'.

   This fact may help us in annotating the multiword phrases in large corpora like the Prague Discourse Treebank, specifically with the annotation of the extent of multiword phrases. In other words, we may better answer the questions like whether to annotate the whole phrases like *s podmínkou, že* 'with the condition that' or only *s podmínkou* 'with the condition' as a connective in examples like Example 1:

[7] Today's conjunction *že* 'that'.

(1) Rodiče mi dovolili koupit si psa **s podmínkou**, **že** úspěšně dodělám školu.

'My parents allowed me to buy a dog **with the condition that** I will successfully finish my school.'

Since we know that *-ž(e)* is a part of many one-word connectives in Czech (from a diachronic point of view), it is very likely also the part of yet non-grammaticalized phrases (that are, at the same time, replaceable by one-word connectives – e.g. the whole *s podmínkou, že* 'with the condition that' in Example 1 is replaceable by one-word *když* 'if', historically also containing the particle *-ž(e)*). In this respect, it may be expected that some of the similar multiword phrases will give rise to a new primary connective in the future, i.e. that *že* 'that' will become part of a new one-word connective as it happened in several cases in the past.

2. The conjunction (former interjection) *a* 'and' is a part of many present-day one-word connectives like ***a**le* 'but', ***a**však* 'however', ***a**č* 'although', ***a**nebo* 'or', ***a**ž* 'untill', ***a**by* 'so that' or unmerged ***a** tak* 'and so', ***a** proto* 'and therefore'. The tendency to combine with *a* 'and' is visible also in present-day multiword phrases (in intra-sentential usage) – see very often phrases like ***a** z tohoto důvodu* 'and for this reason', ***a** to znamená* 'and this means' etc.

3. Another productive formation of connectives is by the negative particle *ne* 'not' – see ***ne**bo* 'or', ***ne**boť* 'for', ***ný**brž*[8] 'but' or ***ne**ž* 'than'.

4. Very frequent is also the combination with the former particle *-le* (with the meaning similar to 'only') – see connectives like *a**le*** 'but', ***le**č* 'however', ***le**da* 'unless'or *a**le**spoň* 'at least'.

5. One of the most productive and also transparent means is the formation of discourse connectives in Czech by combination of prepositions (like *pro* 'for', *přes* 'over', *po* 'after', *za* 'behind', *před* 'before', *při* 'by', *na* 'on, at', *bez* 'without', *v* 'in', *nad* 'over' etc.) and pronouns (especially the demonstrative pronoun *to* 'this' in the whole paradigm) – see one-word examples like *pro**to*** 'therefore', *přes**to*** 'yet, inspite of this', *po**tom*** 'then', *za**tím*** 'meanwhile', *před**tím*** 'before', *při**tom*** 'yet, at the same time', *za**to*** 'however', *na**to*** 'then, after that', *bez**toho*** 'in any case', *v**tom*** 'suddenly', *nad**to***

[8] Originally also ***né**berž(e)*, ***nie**brž*.

'moreover'. Literally, *proto* means 'for this', *přesto* 'in spite of this', *potom* 'after this' etc.

Moreover, there are several present-day prepositional phrases (with discourse connective function) having exactly the same structure like the mentioned one-word connectives (i.e. they consist of a preposition and a demonstrative pronoun *to* 'this'; the only difference is that they have not merged into one-word expression) – see e.g. *kvůli tomu* 'because of this', *navzdory tomu* 'despite this', *kromě toho* 'besides this' etc. signaling discourse relations within a text. Therefore, we consider such prepositional phrases discourse connectives because they express discourse relations within a text and have a similar structure as some one-word connectives – the only difference is that their grammaticalization is not yet completed and that they are not merged into one-word expressions. So it seems that such formation of connectives from prepositional phrases is very productive (not only) in Czech.

A very similar process of discourse connective formation (i.e. from prepositional phrases) may be seen also in other languages, which supports its productivity across languages. The paper demonstrates this on the foreign counterparts of the Czech connective *proto* 'therefore' (that arose from the combination of the preposition *pro* 'for' and pronoun *to* 'this' as mentioned above). English *therefore* arose from the combination of *there* and *fore* (that was an Old English and Middle English collateral form of the preposition *for*) with the meaning 'in consequence of that'. Similar process may be seen in German *dafür* (from the preposition *für* 'for' and deictic component *da*) or parallelly Danish *derfor*. Moreover, there are many other English connectives with similar structure like *thereafter* (meaning 'after that'), *thereupon*, *therein*, *thereby*, *thereof*, *thereto* etc. or in German the productive anaphoric connectives like *davor* 'previously', *danach* 'then' etc. (see Section 4.2). All of these connectives follow the same formation principle (i.e. the anaphoric reference to the previous context plus the given preposition) that seems to be, therefore, language universal. There are similar unmerged phrases in English like *because of this*, *due to this* etc. as potential candidates for grammaticalization, i.e. as potential one-word fixed connectives.

We view the whole structures *because of this*, *due to this* as discourse connectives. As demonstrated above, there are some present-day primary connectives that historically arose from similar combination of a preposition and demonstrative pronoun (e.g. Czech connective *proto* 'therefore' etc.). At the same time, **because of*, **due to* themselves are ungrammatical structures (i.e. we cannot say *The weather is nice. *Due to, I will go to the beach.*) and need to combine with an anaphoric expression to gain a discourse connecting function. For these reasons,

we consider the full structures to be the discourse connectives, i.e. including the demonstrative pronoun *this*.

5.2 Primary connectives and the process of grammaticalization

On the basis of previous analysis, the paper characterizes the most frequent (or prototypical) discourse connectives in the following way.

We use the term PRIMARY CONNECTIVES (firstly introduced by Rysová & Rysová 2014) for expressions with primary connective function (i.e. from part-of-speech membership, they are mainly conjunctions and structuring particles) that are mainly one-word and lexically frozen (from present-day perspective). Primary connectives are synsemantic (or functional) words so they are not integrated into clause structure as sentence elements. The primary connectives mostly do not allow modification (cf. **generally but*, **only and* etc., with some exceptions like *mainly because*). The most crucial aspect of primary connectives is that they underwent the process of grammaticalization, i.e. they arose from other parts of speech (cf., e.g., the connective *too* as the stressed variant of the preposition *to*) or combination of words (cf. English phrases *by cause* ⟶ *because*, *for the reason that* ⟶ *for*, *never the less* ⟶ *nevertheless* etc.), but they merged into a one-word expression during their historical development. Therefore, they underwent the gradual weakening or change of their original lexical meaning and fixing of the new form and function.

At the same time, primary connectives are not a strictly closed class of expressions. They are rather a scale mapping the process of their grammaticalization. This process is sometimes not fully completed so the primary connectives do not have to fulfill all the characteristics stated above – e.g. some of them are still written as two words (like Czech *i když* 'although' or English *as if*, *so that* etc.). The main argument here is that they fulfill most of the aspects and that their primary function in discourse is to connect two pieces of a text.

6 Multiword connecting phrases

6.1 Secondary connectives: Potential candidates for primary connectives?

Apart from primary connectives, also another specific group among discourse connectives may be distinguished – the SECONDARY CONNECTIVES (the term firstly used by Rysová & Rysová 2014). The reason is (as discussed above) that primary

connectives are not the only expressions with the ability to signal discourse relations. There are also multiword phrases like *this is the reason why*, *generally speaking*, *the result is*, *it was caused by*, *this means that* etc. These phrases also express discourse relations within a text (e.g. *generally speaking* signals a relation of generalization), but they significantly differ from primary connectives – mostly, they may be inflected (*for this reason – for these reasons*), modified (*the main/important/only condition is*) and they exhibit a high degree of variation in authentic texts (the variation is better seen in inflected Czech – see, e.g., secondary connectives *příkladem je* vs. *příklad je* both meaning 'the example is', firstly used in instrumental, secondly in nominative). Therefore, secondary connectives may be defined as an open class of expressions.

Generally, secondary connectives are multiword phrases (forming open or fixed collocations) containing an autosemantic (i.e. lexical) component or components. Secondary connectives function as sentence elements (e.g. *due to this*), clause modifiers (*simply speaking*) or even as separate sentences (*the result was clear*). Concerning part-of-speech membership, secondary connectives are a very heterogeneous group of expressions – very often, they contain nouns like *difference*, *reason*, *condition*, *cause*, *exception*, *result*, *consequence*, *conclusion* etc. (i.e. nouns that directly indicate the semantic type of discourse relations), similarly verbs like *to mean*, *to contrast*, *to explain*, *to cause*, *to justify*, *to precede*, *to follow* etc. and prepositions like *due to*, *because of*, *in spite of*, *in addition to*, *unlike*, *on the basis of* (functioning as secondary connectives only in combination with an anaphoric reference to the previous unit of text realized mostly by the pronoun *this* – cf. *due to this*, *because of this* etc.).[9]

All of these aspects indicate that secondary connectives have not yet undergone the process of grammaticalization although they exhibit some of its features – e.g. gradual stabilization or preference of one form or gradual weakening of the original lexical meaning (see Section 6.3).

Within the secondary connectives, the most frequent structures occurring in the PDiT have also been analyzed – see Table 4 (the analysis was done on the annotation of secondary connectives in the PDiT – see Rysová & Rysová 2014; 2015). Table 4 presents the tokens for the individual forms of the secondary connectives, i.e. not lemmas. The aim was to see which concrete form of the same secondary connective is the most frequent and has the biggest chance to become fixed or stable in the future. For example, the PDiT contains the secondary connective *to znamená, že* 'this means that', but also the similar variants like *znamená to, že* [lit.

[9] This type of secondary connectives may be detected in the corpus automatically – see Rysová & Mírovský (2014).

means this that] 'this means that'. In this case, the most frequent is the variant *to znamená, že* 'this means that' with 22 tokens in the PDiT (see Table 4). A high degree of variability is also one of the reasons why secondary connectives are very difficult to annotate in large corpora.

We see that the frequency of the individual secondary connectives is much lower than of the primary connectives (presented in Table 1). The most frequent secondary connective in the PDiT is the verbal phrase *dodal* '(he) added'[10] with 121 tokens. Very frequent secondary connectives are also represented by prepositional phrases (like *v případě, že* 'in case that', *v této souvislosti* 'in this regard'), often in the combination with the demonstrative pronoun *to* 'this' (like *kromě toho* 'besides this' or *naproti tomu* 'in contrast to this'), which is historically a very productive formation of primary connectives (see Section 5.1). One of the most frequent secondary connectives in Czech (in the PDiT) is also the prepositional phrase *z tohoto důvodu* 'for this reason' that is very similar to the Old English phrases such as *for þon þy* literally 'for the (reason) that' giving probably the rise of the present-day English connective *for*.

So it may be observed that the present-day secondary connectives have very similar structures as the former ones and that the process of connective formation thus repeats across the historical development. In very simple terms, the secondary connectives often become primary through the long process of grammaticalization; simultaneously, some new secondary connectives are rising, as well as some old primary connectives are disappearing – cf., e.g., the Old Czech expressions *an*, *ana*, *ano* (lit. 'and he', 'and she', 'and it') being used as connectives for different semantic relations (e.g. conjunction, opposition or reason and result). These expressions were used still in the first half of the 19th century but then they gradually lost their position in language and completely disappeared (see Grepl 1956). In this respect, discourse connectives represent a dynamic complex or set of expressions with stable centre (containing grammaticalized primary connectives) and variable periphery (containing non-grammaticalized secondary connectives).

6.2 Other connecting phrases

During the analysis of the PDiT data, it have been observed that there are also big differences among the multiword connecting phrases themselves – cf. the phrases like *navzdory tomu* 'despite this', *navzdory tomuto faktu* 'despite this fact', *navzdory této situaci* 'despite this situation', *navzdory této myšlence* 'despite

[10] For more details to verbs of saying functioning as secondary connectives see Rysová (2014b).

Table 4: Most frequent secondary connectives in the PDiT

Secondary connectives	Tokens in the PDiT
dodal '(he) added'	121
podobně 'similarly'	60
v případě, že 'in case that'	40
vzhledem k tomu, že 'concerning the fact that'	40
dodává '(he) adds'	36
kromě toho 'besides this'	30
naproti tomu 'in contrast to this'	23
to znamená, že 'this means that'	22
v této souvislosti 'in this regard'	17
případně 'possibly'	13
příkladem je 'the example is'	12
upřesnil '(he) specified'	12
znamená to, že [lit. means this that] 'this means that'	12
z tohoto důvodu 'for this reason'	11

this idea', etc. (all occurring in the authentic Czech texts). All of these phrases clearly signal a discourse relation of concession, but they do not have the same function in structuring of discourse. The difference is that the phrases like *navzdory tomu* 'despite this' may function as discourse connectives in many various contexts (with the relation of concession), i.e. their status of discourse connectives is almost universal or context independent. On the other hand, phrases like *navzdory této myšlence* 'despite this idea' fit only into certain contexts, i.e. they function as indicators of discourse relations only occasionally, not universally (although they contribute to the whole compositional structure of text and participate in text coherence) – see Examples 2 and 3:

(2) Vše začalo nemilým ranním probuzením, všude byla mlha. **Navzdory tomu** jsem sedl do vlaku a odjel.

'Everything started with unpleasant morning awakening, the fog was everywhere. **Despite this**, I sat on the train and left.'

(3) Uvažovali jsme o modernizaci školy a knihovny. **Navzdory této myšlence** došlo z finančních důvodů pouze k rozvoji knihovny.

'We considered modernization of our school and library. **Despite this idea**, we have developed only the library for financial reasons.'

The expression *navzdory tomu* 'despite this' in Example 2 expresses a discourse relation of concession and may be used also in Example 3 (cf. ***Despite this**, we have developed only the library for financial reasons.*). On the other hand, the expression *navzdory této myšlence* 'despite this idea' is more context dependent, i.e. it signals a discourse relation of concession in Example 3 but it cannot be used in Example 2 (cf. *Everything started with unpleasant morning awakening, the fog was everywhere. ****Despite this idea**, I sat on the train and left.*).

This universality (or context independency) is considered a crucial feature of discourse connectives (both primary and secondary) and the boundary between connectives and non-connectives may be put right here, i.e. according to the universality principle.[11] Discourse connectives are thus expressions with (almost) universal connective function, i.e. the author may choose them for signaling given semantic type of discourse relations almost in any context.[12] We do not consider the other phrases (also signaling discourse relations, but only in certain contexts) to be discourse connectives and we call them (non-universal) free connecting phrases.

This paper has tried to demonstrate the heterogeneity of connective means in general (going from grammaticalized primary connectives to variable secondary connectives and free connecting phrases).

6.3 Annotations of discourse connectives and other connecting phrases in large corpora

We believe that the detailed linguistic analysis of discourse connectives and other phrases may help in processing these expressions in large corpora like the Prague Discourse Treebank. As demonstrated above, there are many possibilities to express discourse relations in a language – by one-word, monomorphematic expressions as well as variable multiword phrases. So the annotation in the corpora should react to their variability and different linguistic nature.

At the same time, the annotation of discourse connectives and other connecting phrases in large corpora may significantly help their further examination in terms of how these expressions usually behave in authentic texts.

[11] Universality principle evaluates linguistic expressions from very lexical point of view (i.e. their degree of concreteness and abstractness). It does not reflect, e.g., the differences in register, the degree of subjectivity (cf. the differences between *since* and *because* in English) etc., see Rysová & Rysová 2015.

[12] We are aware that expressions like *and*, *but*, *on the other hand* etc. have also other (non-connective) meanings (cf. *girls and boys*). However, these other meanings are not in our interest – we evaluate the expressions only in their connective function.

The Prague Discourse Treebank contains the annotation of primary connectives (finished in 2012 as PDiT 1.0, see Poláková et al. 2012) and newly also of secondary connectives and other free connecting phrases (published in 2016 as PDiT 2.0, see Rysová et al. 2016); for more information see Rysová & Rysová 2014).[13] Altogether, primary connectives represent 94.6% (20,255 tokens) and secondary connectives 5.4% (1,161 tokens) within all discourse connectives in the PDiT (i.e. altogether 21,416 tokens). So the terms primary and secondary connectives correspond also to their frequency in large corpora. In addition to discourse connectives, the PDiT contains also the annotation of the free connecting phrases (like *despite this idea* etc.) with altogether 151 tokens.

In the current stage, the PDiT thus contains the annotation of explicit discourse relations based on a deep linguistic research, i.e. reflecting all the differences among the individual connective expressions.

The results of the annotation in the PDiT demonstrate that the authors of authentic texts mostly use the grammaticalized primary connectives, then non-grammaticalized secondary connectives and lastly the contextually dependent free connecting phrases. The reasons may be that primary connectives are lexically frozen, short, very often one-word expressions that are not (as functional words) integrated into clause structure. Their usage in texts may thus be related to economy in language, i.e. the author chooses the easiest (or the most economical) solution.

6.4 Secondary connectives in the PDiT vs. alternative lexicalizations of discourse connectives in the PDTB

In the last section, this paper shortly compares the above mentioned approach to discourse connectives in the Prague Discourse Treebank (PDiT) with discourse connectives in the Penn Discourse Treebank (PDTB, see Prasad, Webber & Joshi 2014). The PDTB is one of the richest corpora with discourse annotation and it inspired also the annotation of connectives in the PDiT. Therefore, the paper introduces here where the PDTB and PDiT annotations meet as well as differ with emphasis on multiword discourse phrases (called secondary connectives in the PDiT and alternative lexicalizations of discourse connectives, i.e. AltLexes, in the PDTB).

[13] The inter-annotator agreement on the existence of discourse relations expressed by secondary connectives reached 0.70 F1, agreement of semantic types of relations expressed by secondary connectives is 0.82 (i.e. 0.78 Cohen's κ, see Rysová & Rysová 2015).

The difference in terminology is given by the different approach to discourse connectives in both projects. The terminology reflects especially the annotation strategies of the PDiT and the PDTB that may be briefly described in the following points.

PDTB:

- Explicit connectives (18,459 annotated tokens) – established according to a list of connectives collected from various sources (cf. e.g. Halliday & Hasan 1976; Martin 1992) and updated during the annotations of authentic Wall Street Journal texts; explicit connectives are here restricted to the following syntactic classes: subordinating and coordinating conjunctions, prepositional phrases, adverbs; examples: *so*, *when*, *and*, *while*, *in comparison*, *on the other hand*, *as a result* (see Prasad, Joshi & Webber 2010);
- AltLexes (624 annotated tokens) – discovered during the annotation of implicit relations; the emphasis is placed on the redundancy of AltLexes and explicit connectives in signaling one discourse relation in the same sentence; there are no grammatical restrictions on AltLexes except for they do not belong to explicit connectives – AltLexes are thus viewed as alternatives to explicit connectives; annotation was carried out only between two adjacent sentences; examples: *for one thing*, *one reason is, never mind that*, *adding to that speculation*, *the increase was due mainly to*, *a consequence of their departure could be* (see Prasad, Joshi & Webber 2010).

PDiT:

- Primary connectives (20,255 annotated tokens) – the emphasis is placed on the origin and general characteristics of connectives; primary connectives are mostly grammaticalized synsemantics (grammatical words) without the function of sentence elements; lexically, they are context independent, i.e. they function as primary connectives in many contexts; the annotators were not provided by the list of connectives but acquainted with the general definition; examples: *so*, *when*, *and*, *while*;
- Secondary connectives (1,161 annotated tokens) – they are non-grammaticalized expressions or phrases with the function of sentence elements or sentence modifiers containing lexical (autosemantic) element; lexically, they are context independent, i.e. they function as secondary connectives

in many contexts; they are annotated as a separate group on the whole PDiT data; examples: *in comparison*, *on the other hand*, *as a result*, *for one thing*, *one reason is, never mind that*;

- Other connective means: free connecting phrases (151 annotated tokens) – they are mainly multiword phrases with a high degree of concreteness or lexicality that are highly dependent on context; their annotation is carried out on the whole PDiT data; examples: *adding to that speculation*, *the increase was due mainly to*, *a consequence of their departure could be*.

As we see, both projects look at discourse connectives from slightly different perspective or different point of view, which is reflected both in terminology as well as annotation principles.

7 Conclusion

The paper introduced the analysis of historical formation of discourse connectives especially in Czech. It supports the idea that present-day lexically frozen connectives (called primary) arose from other parts of speech (especially from particles, adverbs and prepositions) or combinations of two or more words. In other words, primary connectives were not primary connectives from their origin but they gained this status during their historical development – through the process of grammaticalization. In this respect, we do not define discourse connectives as a closed class of expressions but rather a scale mapping the grammaticalization of the individual connective expressions.

At the same time, there are two specific groups of discourse connectives: primary and secondary. They differ mainly in the fact in which place on the scale they occur, i.e. whether the process of grammaticalization is already completed (or is in its final phase) or whether this process has just started. In this respect, primary connectives are mainly one-word, lexically frozen, grammatical expressions with primary connecting function and secondary connectives are mainly multiword structures containing lexical (autosemantic) word or words, functioning as sentence elements, clause modifiers or even separate sentences. Both primary and secondary connectives are defined on the basis of their context independency (i.e. on their suitability to function as connectives for given semantic relation in many various contexts).

Since the present-day primary connectives arose from similar phrases or parts of speech like secondary connectives (and very often from combination of several

words that gradually merged together – with some possible losses), we look at the secondary connectives as at the potential primary connectives in the future.

The paper has also analyzed another group of connective expressions – the free connecting phrases (like *despite this idea*, *because of these activities* etc.) functioning as discourse indicators only occasionally, depending on certain contexts, i.e. these phrases do not have a universal status of discourse connectives (as both primary and secondary) and they exhibit a high degree of variation.

The paper has shown the etymology and historical origin of the most frequent discourse connectives especially in Czech, English and German. It was found out that the examined connectives exhibit a similar behaviour and that they underwent a similar process of formation. In this respect, tha paper suggests that the rise and ways of formation of discourse connectives is (to large extent) language universal.

The analysis may help with the annotation of discourse in large corpora, as the annotation principles should react to the differences among the individual connective expressions and should be based on a detailed theoretical research. We have carried out such annotation in the Prague Discourse Treebank (on almost 50 thousand sentences) to observe how these expressions behave in authentic texts and what is their frequency in the large corpus data. We found out that primary connectives represent 94.6% and secondary connectives 5.4% within all discourse connectives in the PDiT. The most frequent secondary connectives have very similar structures that gave rise to present-day primary connectives.

Acknowledgments

The author acknowledges support from the Czech Science Foundation (Grant Agency of the Czech Republic): project GA CR No. 17-06123S (Anaphoricity in Connectives: Lexical Description and Bilingual Corpus Analysis). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

The author gratefully thanks Jiří Mírovský from the Charles University for providing quantitative data on the basis of the PDiT for this paper.

References

Aijmer, Karin. 2002. *English discourse particles: Evidence from a corpus*. Vol. 10. Amsterdam: John Benjamins Publishing.

Bauer, Jaroslav. 1962. Spojky a příslovce [Conjunctions and adverbs]. *Sborník prací FF BU, A10* 11. 29–37.

Bauer, Jaroslav. 1963. Podíl citoslovcí na vzniku českých spojek [The importance of interjections in the development of Czech conjunctions]. *Sborník prací FF BU, A* 11. 21–28.

Claridge, Claudia. 2013. The evolution of three pragmatic markers: As it were, so to speak/say and if you like. *Journal of Historical Pragmatics* 14(2). 161–184.

Claridge, Claudia & Leslie Arnovick. 2010. Pragmaticalisation and discursisation. *Historical Pragmatics* 8. 165–169.

Cuenca, Maria-Josep. 2003. Two ways to reformulate: A contrastive analysis of reformulation markers. *Journal of Pragmatics* 35(7). 1069–1093.

Degand, Liesbeth & Jacqueline Evers-Vermeul. 2015. Grammaticalization or pragmaticalization of discourse markers?: More than a terminological issue. *Journal of Historical Pragmatics* 16(1). 59–85.

Degand, Liesbeth & Anne-Marie Simon Vandenbergen. 2011. Introduction: Grammaticalization and (inter) subjectification of discourse markers. *Linguistics* 49(2). 287–294.

Fischer, Kerstin. 2006. *Approaches to discourse particles*. Amsterdam: Elsevier.

Fraser, Bruce. 1999. What are discourse markers? *Journal of Pragmatics* 31(7). 931–952.

Grepl, Miroslav. 1956. Spojka an… Ve spisovném jazyce první poloviny 19. Století. *Sborník prací Filosofické fakulty brněnské university A 4* 5. 45–50.

Hajičová, Eva, Barbara Partee & Petr Sgall. 2013. *Topic-focus articulation, tripartite structures, and semantic content*. Springer Science & Business Media.

Hakulinen, Auli. 1998. The use of Finnish nyt as a discourse particle. *Pragmatics and Beyond New Series* 57. 83–96.

Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Hansen, Maj-Britt Mosegaard. 1998. *The function of discourse particles: A study with special reference to spoken standard French*. Vol. 53. Amsterdam: John Benjamins Publishing.

Harper, Douglas et al. 2001. *Online etymology dictionary*. http://etymonline.com/.

Holub, Josef & František Kopečný. 1952. *Etymologický slovník jazyka českého [Etymological dictionary of Czech]*. Prague: Státní nakladatelství učebnic v Praze.

Klein, Wolfgang & Alexander Geyken. 2010. Das Digitale Wörterbuch der Deutschen Sprache (DWDS). *Lexicographica* 26. 79–93.

Lenker, Ursula & Anneli Meurman-Solin. 2007. *Connectives in the history of English*. Amsterdam: John Benjamins Publishing.

Martin, James R. 1992. *English text: System and structure*. Amsterdam: John Benjamins Publishing.

Maschler, Yael. 2000. *Discourse markers in bilingual conversation*. Kingston Press Services.

Poláková, Lucie, Pavlína Jínová, Šárka Zikánová, Eva Hajičová, Jiří Mírovský, Anna Nedoluzhko, Magdaléna Rysová, Veronika Pavlíková, Jana Zdeňková, Jiří Pergler & Radek Ocelák. 2012. *Prague Discourse Treebank 1.0*. Prague, Czech Republic: ÚFAL MFF UK.

Prasad, Rashmi, Aravind K. Joshi & Bonnie Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 1023–1031. Association for Computational Linguistics.

Prasad, Rashmi, Bonnie Webber & Aravind K. Joshi. 2014. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics* 40(4). 921–950.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi & Bonnie L. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*.

Rejzek, Jiří. 2001. *Český etymologický slovník [Czech etymological dictionary]*. nakladatelství LEDA.

Rysová, Kateřina. 2014a. *O slovosledu z komunikačního pohledu [On word order from the communicative point of view]* (Studies in Computational and Theoretical Linguistics). Prague: ÚFAL.

Rysová, Magdaléna, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková & Šárka Zikánová. 2016. *Prague Discourse Treebank 2.0*. Prague, Czech Republic: ÚFAL MFF UK.

Rysová, Magdaléna. 2012. Alternative lexicalizations of discourse connectives in Czech. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2800–2807. Istanbul, Turkey: European Language Resources Association.

Rysová, Magdaléna. 2014b. Verbs of saying with a textual connecting function in the Prague Discourse Treebank. In *Proceedings of the Ninth International Con-*

ference on Language Resources and Evaluation (LREC 2014), 930–935. Reykjavik, Island: European Language Resources Association.

Rysová, Magdaléna & Jiří Mírovský. 2014. Use of coreference in automatic searching for multiword discourse markers in the Prague Dependency Treebank. In Lori Levin & Manfred Stede (eds.), *Proceedings of The 8th Linguistic Annotation Workshop (LAW-VIII)*, 11–19. Dublin City University (DCU). Dublin, Ireland: Dublin City University (DCU).

Rysová, Magdaléna & Kateřina Rysová. 2014. The centre and periphery of discourse connectives. In Wirote Aroonmanakun, Prachya Boonkwan & Thepchai Supnithi (eds.), *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, 452–459. Department of Linguistics, Faculty of Arts, Chulalongkorn University. Bangkok, Thailand: Department of Linguistics, Faculty of Arts, Chulalongkorn University.

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

Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Schourup, Lawrence. 1999. Discourse markers. *Lingua* 107(3). 227–265.

Shloush, Shelley. 1998. A unified account of Hebrew bekicur 'in short': Relevance theory and discourse structure considerations. *Discourse Markers: Descriptions and Theory* 57. 61–82.

Urgelles-Coll, Miriam. 2010. *The syntax and semantics of discourse markers*. London: Continuum International.

Zwicky, Arnold M. 1985. Clitics and particles. *Language* 61(2). 283–305.