THE CORRELATION BETWEEN DISCOURSE-ANAPHORIC DEVICES AND AN OVERALL COMMUNICATIVE COMPETENCE IN LEARNERS' ESSAYS

Kateřina Rysová, Magdaléna Rysová

Charles University, Faculty of Mathematics and Physics (CZECH REPUBLIC)

Abstract

In the present paper, we introduce a contrastive study on text coherence in Czech and German. Specifically, we focus on the discourse-anaphoric devices (called anaphoric connectives) contributing to text coherence and we analyze their role in the overall communicative competence of both native and non-native speakers of Czech and German. In our analysis, we firstly examine anaphoric connectives in Czech and we present their frequencies in three corpora – SYN6 and PDiT 2.0 (for texts by native speakers), and MERLIN (for texts by non-native speakers). Then we take a closer look at anaphoric connectives in texts by non-native speakers (learners) of Czech. Specifically, we examine whether the students who actively use such expressions in their writings also have better communicative competence as a whole (i.e. they reach better overall grade). Subsequently, we focus on the ways anaphoric connectives in Czech are translated into German (using the corpus InterCorp) and we provide an analysis of these German counterparts. In the next step, we carry out the same type of analysis for anaphoric connectives in German (with the PCC and the DWDS corpora for native speakers and the MERLIN corpus for non-native speakers). We analyze the results for Czech and German separately, and finally, we carry out a comparative study of these two languages with a focus on the use of our findings in the teaching process and possibly in (automated) translation.

Keywords: Communicative competence, discourse, anaphora, anaphoric connectives, discourse connectives, Czech as a foreign language, German as a foreign language.

1 INTRODUCTION

In this paper, we focus on the analysis of text coherence, i.e. on the fluency and continuity of text. The fact that the well-formulated text is easily comprehensible for the reader is caused by various linguistic and extra-linguistic factors – e.g. by the thematic arrangement of the text, the logical continuity of the presented information, the formal structuring into paragraphs, or the appropriate use of language means (discourse connectives) expressing the relations between text units. It is natural in a language that some of these factors even combine and as such, they should be given a special attention when studying complex phenomena as text coherence.

In this study, we examine discourse-anaphoric devices as essential representatives of the general group of cohesive devices that help to constitute the well-structured and coherent text. These devices (called also anaphoric connectives) are expressions such as *therefore, thereby, thereof* (in Czech, e.g., *proto, přesto, potom* or in German *darum, trotzdem, danach*). The use of an anaphoric connective in the text is illustrated in Example (1) presenting a short part of Tolkien's *The Silmarillion* and its German and Czech translations (anaphoric connectives are in bold).

(1) English: *The white timbers we wrought with our own hands, and the white sails were woven by our wives and daughters.* ***Therefore*** *we will neither give them nor sell them for any league or friendship.*

German: *Ihre weißen Planken haben wir mit eigenen Händen gezimmert, und die weißen Segel haben unsere Frauen und Töchter gewoben.* ***Darum*** *geben wir sie nicht her, und sie sind uns nicht feil, für keinen Bund und keine Freundschaft.*

Czech: *Bílé kameny jsme opracovávali vlastníma rukama a bílé plachty tkaly naše manželky a dcery.* ***Proto*** *je ani nedáme, ani neprodáme pro žádné spojenectví ani přátelství.*

All three expressions in bold are originally combinations of a preposition (*for(e)* in English, *um* in German and *pro* in Czech) and an anaphoric expression (*there* in English, *da* in German and *to* in Czech). Anaphoric connectives are special in the way that they combine two discourse phenomena. They connect two text units and signal a semantic relation between them (e.g. *therefore* expresses a relation of result), similarly as other "non-anaphoric" connectives (e.g. *and, but, while* or *because*). At

ISBN: 978-84-09-02709-5

the same time, however, these expressions contain an inherent (internal) anaphoric reference, e.g. *there*, and they constitute both discourse and referential net of relations as the two essential pillars of text coherence. The discourse-anaphoric devices are thus complex expressions having a substantial role in text production and their correct using helps to improve and deepen the author's communicative skills.

In this paper, we carry out research of anaphoric connectives originally consisting of a preposition and an anaphoric expression (cf. *therefore, darum, proto*), i.e. of originally multi-word anaphoric units that are currently lexicalized as one-word connectives. The aim of our work is to analyze these expressions contrastively: i) on two typologically different languages: Czech and German, ii) as well as on two different text types: texts written by native speakers and language learners.

Generally, discourse connectives are lexical anchors of semantico-pragmatic text relations (e.g. *therefore*, *because*, *but*, *for this reason*, *on the other hand*, *in summary*) that may further be divided into two groups – primary (e.g. *therefore*) and secondary (e.g. *for that reason*), differing in the degree of grammaticalization as defined in [15] and described in detail in [16]). In this respect, we examine selected primary connectives, i.e. connectives that are lexicalized as one-word expressions. Anaphoric connectives form then a special subgroup of connectives and they represent a relatively new research topic. Division of connectives into structural (taking arguments qua the syntactic configuration) and anaphoric (mostly adverbials) is given in [24], based on the English material. Anaphoric connectives in German are studied in [20]. For Czech, the first studies of anaphoric connectives consisting of a preposition and a demonstrative pronoun are given in [9] and in [14]. The German-English-Czech comparison of anaphoric connectives can be found at [7].

2 METHODOLOGY

First, we search for typical representatives of one-word anaphoric connectives in Czech (Section 4.1) and German (Section 4.2), originally consisting of prepositional phrases. For this purpose, we use the Prague Discourse Treebank 2.0 (PDiT) [18] for Czech which contains a detailed annotation of discourse connectives (covering the anaphoric ones), and the German Discourse Marker Lexicon (DimLex) [19, 22] aiming to cover all German discourse connectives in current use. First, we examine the frequency of the anaphoric connectives in texts written by native speakers of Czech and German. For Czech, we use the Prague Discourse Treebank 2.0 and the Czech National Corpus (CNC, version SYN6) [6] (Section 4.1.1). For German, we work with the Potsdam Commentary Corpus (PCC) [21] and with the DWDS-Kernkorpus (version 3) [4] (Section 4.2.1).

Subsequently, we focus on the anaphoric connectives in the communication of non-native speakers of Czech (Section 4.1.2) and German (Section 4.2.2). We examine the frequency of selected anaphoric connectives in texts written by language learners reaching different degrees of language proficiency. Their communicative competence is assessed in accordance with the Common European Framework of Reference for Languages (CEFR) [2]. We analyze texts written by learners at A2, B1, and B2 levels that means on the levels corresponding to basic language users and independent language users. For studying anaphoric connectives in the non-native speakers' data, we use a multi-language corpus MERLIN [1] which contains both Czech and German texts written by language learners.

In the next step, we find the German counterparts of the most frequent anaphoric connectives in Czech (Section 4.1.3) and vice versa (Section 4.2.3) using the tool Treq [23] working with the parallel Czech and German language data. Finally, we compare the Czech-German and German-Czech counterparts of anaphoric connectives in general (using the Treq tool) and we provide general conclusions on them from the perspective of the teaching process and (automated) translation (Section 4.3).

3 LANGUAGE DATA

The **Prague Discourse Treebank** (PDiT) is a corpus built on the data of the Prague Dependency Treebank [5]. It consists of Czech newspaper texts and contains 3,165 annotated documents (49,431 sentences and 833,195 tokens). The PDiT contains a detailed annotation of semantico-pragmatic text relations expressed by discourse connectives (including the anaphoric ones). The first version of the PDiT was published as [10] (described in detail in [11]), the second version as [18]. For our analysis, we use the second one, PDiT 2.0 (reflecting the division of connectives into primary and secondary [15]). **Czech National Corpus** is a large collection of various corpora. We deal with the corpus SYN6 [6] containing 15,494,077 texts (307,694,879 sentences, 4,834,739,998 tokens). It predominantly consists of newspaper texts automatically annotated with morphology and syntax.[1] For the comparative analysis, we use the parallel corpus InterCorp, see [3] and [13]. It covers about 40 languages, among others also Czech and German. The German part (parallel to the Czech one) contains 6,543,622 sentences (101,360,040 positions). It consists of journalistic, legal and administrative texts, Bible translations, subtitles, and fiction. **Potsdam Commentary Corpus** [21] contains 220 German newspaper texts (2,900 sentences, 44,000 tokens). It is annotated with sentence syntax, coreference, discourse structure, and connectives. The discourse annotation follows the concept of the Penn Discourse Treebank [12] (with minor modifications). The delimitation of connectives is based on the concept described in [8]. **DWDS-Kernkorpus** [4] is a corpus of German texts by native speakers. It consists of fiction, newspaper texts, science, and utility literature. It contains 79,116 documents (5,819,576 sentences, 121,397,604 tokens). It is annotated with lemmas and parts of speech. **German Discourse Marker Lexicon** [19, 22] (DimLex) is a lexicon covering 275 connectives currently used in German, both anaphoric and non-anaphoric. The current version aims to cover all known German discourse connectives. The **MERLIN** corpus [1] contains 1,462 texts written by learners of German and Czech. The texts are marked with the CEFR level and annotated with lemmas and parts of speech as well as with error annotation. The German part contains 1,024 texts (A1: 57 texts, A2: 297 texts, B1: 331 texts, B2: 293 texts, C1: 42 texts, and C2: 4 texts), the Czech part consists of 438 texts (A1: 1 text, A2: 189 texts, B1: 165 texts, B2: 81 texts, and C1: 2 texts).

4 RESULTS

4.1 Anaphoric connectives in Czech

4.1.1 Frequencies in Czech corpora

We have extracted anaphoric connectives in Czech from the PDiT 2.0. The primary connectives in form of grammaticalized structures consisting of a preposition and a demonstrative pronoun build a closed class of expressions. There are about 15 connectives of this type in Czech, see Table 1. The individual Czech anaphoric connectives do not occur in the language with the same frequency. Table 1 demonstrates how often these Czech connectives occur in the data of the Czech National Corpus, the Prague Discourse Treebank (both containing texts written by native speakers of Czech) and in the MERLIN (corpus of non-native speakers' texts).

Table 1. *Czech anaphoric connectives in corpora of texts by native and non-native speakers*

Czech grammaticalized anaphoric connectives	CNC SYN6 (native speakers)	PDiT 2.0 (native speakers)	MERLIN (non-native speakers)
protože "because"	3,207,195	640	259
proto "therefore"	3,190,403	487	90
přitom "while"	1,438,906	261	4
poté "afterwards"	1,253,348	73	1
přesto "yet"	1,131,699	158	2
zatímco "while"	906,394	207	3
potom "then"	653,810	96	37
přestože "though"	564,489	124	4
předtím "before"	381,392	66	3
přičemž while"	293,306	92	0
zato "but still"	166,387	46	0
mezitím “meanwhile"	135,605	32	0
nato "thereafter"	83,026	7	0
natož "let alone"	49,081	14	0
nadto "moreover"	9,981	3	0
mimoto "besides"	7,330	5	0

[1] Most of the used corpora are annotated automatically and their annotation cannot be taken as 100% reliable. At the same time, most of them do not contain discourse annotation (i.e. annotation of discourse connectives) – since most of our data are also processed automatically, we cannot easily distinguish connective and non-connective usages in case of polysemy. The exact numbers in the following analysis should thus be taken rather as tendencies concerning Czech and German connectives (and their translations).

Table 1 shows that the first two most frequent grammaticalized anaphoric connectives in Czech are the same ones in all used corpora. The most frequent connective is *protože* "because", the second one is *proto* "therefore". According to our data, the high frequency of *protože* and *proto* in written language is shared by native speakers as well as learners of Czech.

The corpora containing texts by native speakers (CNC, PDiT) share also the third most frequent connective, namely *přitom* "while". On the contrary, *přitom* is not frequently used by non-native speakers. They prefer the connective *potom* "then" as the third most frequent one in the MERLIN corpus. The CNC and PDiT agree also on the tendency of usage of the less frequent anaphoric connectives. The less frequent ones are mainly *zato, mezitím, nato, natož, nadto*, and *mimoto* in both corpora. Generally, we can say that the order of grammaticalized anaphoric connectives according to their frequencies is very similar in the CNC and PDiT.

4.1.2 Czech anaphoric connectives in learners' texts

In this section, we focus on anaphoric connectives in texts by non-native speakers of Czech and we observe whether there is some development in their use across the individual CEFR levels. We select the three most frequent Czech anaphoric connectives in MERLIN: *protože* "because", *proto* "therefore", *potom* "then" (other anaphoric connectives have very low frequency), see Table 2. In our study, three groups of text levels are distinguished: A2, B1, B2.[2]

Table 2. *Most frequent anaphoric connectives in Czech in learners' texts from MERLIN corpus*

CEFR level of the text	Occurrences of ***protože*** per 100 texts	Occurrences of ***proto*** per 100 texts	Occurrences of ***potom*** per 100 texts
A2	57	12	6
B1	63	19	13
B2	54	41	5

Table 2 shows that each connective occurs with similar frequency at the CEFR levels. Specifically, there is no significant difference in the use of *protože* in A2-B1-B2 levels and in the use of *potom* in A2-B1-B2 levels (measured by chi-square test, p-value ≤ 0.001). The only significant difference is in the occurrence of the connective *proto* (a significant difference is between levels A2 and B2, as well as between levels B1 and B2). However, generally, the frequency of the anaphoric connectives in the texts by non-native speakers of Czech is rather low, which demonstrates that reaching coherence through these devices is probably difficult for the learners (or not yet fully adopted by them). This corresponds to the similar findings on Czech connective expressions presented in [17].

4.1.3 Czech anaphoric connectives and their German counterparts

As mentioned in Section 4.1.1, there are about 15 grammaticalized anaphoric connectives in Czech originating from the combination of a preposition and an anaphoric expression. Table 3 presents also their most frequent German counterparts (according to the InterCorp data).

Table 3 demonstrates that the German counterparts of Czech grammaticalized anaphoric connectives may be both anaphoric (e.g. *außerdem, überdies, darauf, daher, trotzdem, dabei*) and non-anaphoric (e.g. *inzwischen, geschweige, weil, aber*).[3] A closer look at the German counterparts indicates that German has much more grammaticalized anaphoric connectives at its disposal than Czech.

In Table 3, there are usually several grammaticalized anaphoric connectives in German corresponding to a single one in Czech. For example, the Czech connective *proto* "therefore" has five very frequent German grammaticalized anaphoric synonyms (according to the corpus InterCorp): *daher, deshalb, darum, deswegen, demnach*. Similarly, *nadto* "moreover" has *überdies, außerdem,* and *zudem* as its most frequent equivalents. Therefore, German tends to combine discourse and anaphoric phenomena and to create grammaticalized anaphoric connectives much more than Czech. We check this hypothesis in Section 4.2.

[2] Other CEFR levels for Czech are represented by a small number of texts and they are thus not included in the analysis.

[3] However, we need to emphasize that the synonyms captured in Table 3 are contextual synonyms. It means they work as synonyms in some contexts but not necessarily in all the possible ones, for details see Section 4.3.

Table 3. *Czech grammaticalized anaphoric connectives and their most often German counterparts*

Czech grammaticalized anaphoric connectives	German equivalents (occurrences in InterCorp)
mezitím "meanwhile"	inzwischen (913), mittlerweile (255), unterdessen (244), indessen (94), in der Zwischenzeit (79), dazwischen (70), währenddessen (48), zwischendurch (36), zwischenzeitlich (27)
mimoto "besides"	außerdem (188), darüber hinaus (47), überdies (19)
nadto "moreover"	überdies (69), außerdem (63), zudem (57), darüber hinaus (53), obendrein (22), zumal (14)
nato "thereafter"	darauf (703), später (485), danach (160), daraufhin (160)
natož "let alone"	geschweige (90), gar (64), ganz zu schweigen von (18)
potom "then"	dann (10,459), danach (939), nachher (309), darauf (284), nachdem (275), anschließend (127), hinterher (116), sodann (61)
poté "afterwards"	nachdem (1,646), dann (952), danach (616), darauf (115), daraufhin (62), sodann (43), nachher (17)
proto "therefore"	daher (9,928), deshalb (9,394), darum (2,642), aus diesem Grund (1,964), deswegen (1,582), folglich (250), demnach (119)
protože "because"	weil (20,523), denn (14,842), da (12,262)
předtím "before"	zuvor (1,254), vorher (1,170), bevor (735), früher (300), vorhin (265), davor (93), vordem (41)
přesto "yet"	trotzdem (2,146), dennoch (2,045), doch (1,919), jedoch (366)
přestože "though"	obwohl (1,922), obgleich (267), zwar (217), wenngleich (77), trotzdem (56), obschon (35)
přitom "while"	dabei (3,548), zugleich (93), hierbei (40)
přičemž "while"	wobei (803), während (311), dabei (207)
zato "but still"	dafür (589), aber (213), dagegen (75), hingegen (50)
zatímco "while"	während (6,886), obwohl (149), wohingegen (88), indes (51), indessen (46)

4.2 Anaphoric connectives in German

4.2.1 Frequencies in German corpora

Based on DimLex, we selected grammaticalized anaphoric connectives in German (coming from the combination of a preposition and an anaphoric expression). Then we searched for their frequencies in German corpora (containing texts by native as well as non-native speakers as in the case of Czech). However, we face here an issue of ambiguity of these expressions. Not all the selected German lexemes occur only in the role of discourse connectives, see the expression *darauf* as a connective in Example (2) and in a non-connective function in Example (3).

(2) *Da wichen die Orks und flohen, und die Eldar hatten den Sieg, und ihre berittenen Bogenschützen verfolgten die Feinde bis in die Eisenberge.* ***Darauf*** *regierte Húrin, Galdors Sohn, über das Volk Hadors in Dor-lómin und diente Fingon.*

“Then the Orcs broke and fled, and the Eldar had the victory, and their horsed archers pursued them even into the Iron Mountains. **Thereafter** Húrin son of Galdor ruled the house of Hador in Dor-lómin, and served Fingon.”

(3) *Ich freue mich sehr* ***darauf****.*

“I’m looking forward **to it** very much.”

Not all the available corpora we work with contain a discourse annotation that is able to distinguish between connective and non-connective uses. Therefore, we analyze the selected expressions as lexemes in Table 4 and in Table 5.

Although some of the expressions in Table 4 may be ambiguous, i.e. they may be used both in a connective and non-connective function (see Examples (2) and (3) above), it is not true for all of them. Some forms (e.g. *demzufolge*) are always (or at least mostly) connective. The connective and non-connective function of German lexemes is distinguished by *Wörterbuch der Konnektoren* [25]. Therefore, we searched for all the expressions listed in Table 4 in this dictionary – those in which the dictionary does not mention their non-connective uses are put in bold. We can assume that these expressions occur almost exclusively or at least predominantly as discourse connectives.

We see from the table that German has more than 50 grammaticalized anaphoric connectives (based on DiMLex), which is roughly 3 times more than in Czech. It is thus true that occurrence of these types of connectives is more typical for German and it supports the general notion of vocabulary differences between these two languages: German tends to compounding words to a greater extent than Czech.

Table 4. *German anaphoric expressions in corpora of texts by native and non-native speakers*

Lexeme	DWDS	PCC	MERLIN	Lexeme	DWDS	PCC	MERLIN	Lexeme	DWDS	PCC	MERLIN
damit	70,788	45	55	**nachher**	4,806	0	2	**hiermit**	1,620	0	16
dabei	47,321	20	14	worauf	4,396	0	2	**demgegenüber**	1,601	0	0
dazu	43,693	16	38	**seitdem**	4,309	1	2	**hierdurch**	1,283	0	0
darauf	43,266	12	29	**deswegen**	3,721	1	29	**stattdessen**	1,190	0	1
dafür	29,092	19	50	**hingegen**	3,639	2	3	**mithin**	1,189	1	0
daher	25,460	3	11	wonach	3,497	1	0	**unterdessen**	865	0	0
dadurch	23,319	2	6	wodurch	3,427	0	0	wogegen	724	0	0
indem	19,712	4	0	**zudem**	3,209	4	0	**demzufolge**	611	0	0
dagegen	19,468	7	4	womit	3,042	0	1	hiernach	482	0	0
nachdem	18,835	4	11	davor	2,872	0	1	**nebenher**	425	0	0
darum	15,960	4	11	**daraufhin**	2,783	1	1	weswegen	349	0	0
wobei	15,336	1	0	**hierfür**	2,780	0	0	**währenddessen**	337	0	0
danach	11,936	1	19	**demnach**	2,612	0	0	**woraufhin**	199	0	0
vorher	11,659	2	5	**überdies**	2,560	0	0	**wohingegen**	138	0	0
trotzdem	10,153	3	19	**infolgedessen**	2,237	0	0	**dahingegen**	35	0	0
hierzu	7,558	0	0	**hierauf**	2,086	0	0	**dementgegen**	2	0	0
indessen	6,193	1	0	**hinterher**	1,853	0	0	**hieraufhin**	1	0	0
daneben	5,170	0	0	**seither**	1,782	1	1				

Table 4 shows that in all three corpora, the most common lexemes are *damit* lit. "with it", *dazu* lit. "to it", *darauf* lit. "on / about / after that" and *dafür* lit. "for it". The absolutely most common anaphoric expression is *damit*. The order of other frequent lexemes slightly differs across the corpora. Here are the first five most frequent lexemes in each corpus.

DWDS-Kernkorpus: *damit, dabei, dazu, darauf, dafür*; PCC: *damit, dabei, dafür, dazu, darauf*; MERLIN: *damit, dafür, dazu, darauf, deswegen*.

A similar observation was reached also in the Czech data: the most frequent connectives are the same ones in corpora of texts written by native as well as non-native speakers. We can thus conclude that language learners acquire the way of using these expressions (whether intentionally or unintentionally) mostly in accordance with the habits of native speakers.

4.2.2 German anaphoric expressions in learners' texts

In the next step, we examine using of anaphoric expressions in German by non-native speakers. Again, we select the most frequent ones in corpus MERLIN (as in case of Czech, we choose those having at least 30 occurrences) and we observe their relative frequencies across the A2–B2 levels according to CEFR.

Table 5. *Most frequent anaphoric expressions in German in learners' texts from MERLIN corpus*

CEFR level of the text	**Occurrences of *damit* per 100 texts**	**Occurrences of *dafür* per 100 texts**	**Occurrences of *dazu* per 100 texts**
A2	0	0	0
B1	10	9	6
B2	11	11	9

Table 5 demonstrates that the selected anaphoric expressions do not occur at all at A2 level. On the other hand, they are used at levels B1 and B2 – interestingly with a similar frequency (there is no significant difference between the occurrence of *damit* in B1 and B2, *dafür* in B1 and B2 and *dazu* in B1 and B2; measured by chi-square test, p-value ≤ 0.001). To conclude, there is a clear jump between the levels A and B, i.e. between basic and independent language users.

4.2.3 German anaphoric connectives and their Czech counterparts

When studying the use of anaphoric connectives in Czech and German and their possible counterparts, it is necessary to examine both ways of translations, i.e. translations from Czech to German (as analyzed in Section 4.1.3) and from German to Czech (analyzed here). In the next step, we thus search for the most frequent anaphoric connectives (based on Table 4) in the original German texts and we examine their counterparts in Czech translation. We use the parallel corpus InterCorp.[4]

Table 6. *German grammaticalized anaphoric connectives and their Czech counterparts*

German lexemes	Czech equivalents (occurrences in InterCorp)
damit “so (that)”	aby (14,749), to (5,524), tedy (545), což (459)
dabei “at the same time / yet”	přitom (4,085), přičemž (98)
dazu “in addition”	navíc (223)
darauf “thereafter”	nato (642), potom (116), poté (115), pak (93)
dafür “but still”	zato (589)
daher “therefore”	proto (15,714), tedy (3,591), tudíž (1,321), takže (437), tak (340)
dadurch “so”	ten (1,809), tak (314), proto (108)
indem “so”	tím, že (3,313), neboť (163), protože (88), přičemž (57)
dagegen “on the other hand”	proti tomu (1,389), naproti tomu (419), naopak (302), však (125), zato (75)
nachdem “thereafter”	když (2,247), poté (1,646), jakmile (218), potom (87)
darum “therefore”	proto (1,939), tak (76), tedy (53), takže (39), tudíž (29)
wobei “while”	přičemž (1,995), přitom (159)
danach “thereafter”	pak (958), potom (768), poté (616), nato (128)
vorher “before”	předtím (1,107), dřív(e) (527), napřed (43)
trotzdem “in spite of this”	přesto (2,201), nicméně (504), stejně (255), však (230), ale (197), přece (185)
hierzu “for this purpose”	za tím(to) účelem (201), v této souvislosti (44), vzhledem k tomu (15), proto (12)

It is noticeable that Czech counterparts of German anaphoric connectives are not as diverse as German counterparts of Czech anaphoric connectives (see Table 3). The fact that German tends to compounding connectives much more than Czech is supported also by the use of non-grammaticalized secondary connectives as Czech equivalents. For example, one-word connective *dagegen* in German is mostly translated as *proti tomu* and *naproti tomu* in Czech, both these counterparts being non-grammaticalized prepositional phrases (called secondary connectives according to [15]). Similarly, Czech prepositional phrases *za tím(to) účelem*, *v této souvislosti* and *vzhledem k tomu* are three most frequent counterparts to the one-word connective *hierzu* in German.

4.3 Anaphoric connectives in Czech and German – general findings

In this part, we draw the general conclusions on Czech and German anaphoric connectives resulting from our analysis presented above. We especially pay a closer attention to the Czech-German and German-Czech translations and to the correspondence between the individual connectives in these two languages which could be used in automated translation.

The proximity and distance between anaphoric connective in Czech and their German equivalents (or better to say the degree of equivalency of the individual pairs of connectives) is measured in Table 7.

[4] The Czech counterparts were firstly found automatically using the Treq tool. Then we distinguished the connective and non-connective usages, according to the detected Czech counterparts – Table 6 should capture mainly the connective functions.

Table 7. *Czech-German vs. German-Czech counterparts of discourse connectives (all are grammaticalized anaphoric connectives in Czech)*

Czech-German counterparts	Percentage of occurrences according to Treq	**Czech-German counterparts**	Percentage of occurrences according to Treq
mezitím – inzwischen "meanwhile"	45%	inzwischen – mezitím	39%
mimoto – außerdem "besides"	57%	außerdem – mimoto	3%
nadto – überdies, außerdem, zudem, darüber hinaus "moreover"	17–18%	überdies – nadto (außerdem – nadto) (zudem – nadto) (darüber hinaus – nadto)	9% (1%) (2%) (1%)
nato – darauf "thereafter"	37%	darauf – nato	10%
natož – geschweige "let alone"	61%	geschweige – natož	67%
potom – dann "then"	78%	dann – potom	18%
poté – nachdem "afterwards"	32%	nachdem – poté	26%
proto – daher "therefore"	37%	daher – proto	65%
protože – weil "because"	42%	weil – protože	61%
předtím – zuvor "before"	34%	zuvor – předtím	40%
přesto – trotzdem, dennoch "yet"	29–30%	trotzdem – přesto (dennoch – přesto)	54% (39%)
přestože – obwohl "though"	57%	obwohl – přestože	22%
přitom – dabei "while"	73%	dabei – přitom	45%
přičemž – wobei "while"	60%	wobei – přičemž	52%
zato – dafür "but still"	50%	dafür – zato	8%
zatímco – während "while"	89%	während – zatímco	34%

Table 7 demonstrates that half of the listed connectives have the same counterpart from the viewpoint from Czech to German and also from German to Czech. For example, the most frequent counterpart of the Czech connective *mezitím* "meanwhile" is the German connective *inzwischen* "meanwhile" (the word *mezitím* was translated as *inzwischen* in 45% of all its occurrences), and also the Czech *mezitím* is the most frequent equivalent for the German *inzwischen* (the word *inzwischen* was translated as *mezitím* in 39% of all its occurrences). Such cases are tinged in grey in Table 7.

In the second half of the listed connectives (not in grey), the Czech-German most frequent equivalents are not the most frequent German-Czech equivalents. For example, the German connective *außerdem* is the most frequent counterpart for the Czech connective *mimoto* "besides". However, the first counterpart of the German *außerdem* is the Czech multiword anaphoric connective *kromě toho* "besides" (*mimoto* is the 6th one). Similarly, the German *obwohl* is the first equivalent for the Czech *přestože* "though"; however, the most frequent counterpart of *obwohl* is *ačkoli* in Czech.

Table 7 also shows that some connectives have more foreign synonyms that seem to be equally relevant. For example, the Czech connective *nadto* "moreover" has four German counterparts (*überdies, außerdem, zudem, darüber hinaus*) occurring with very similar frequency; similarly, the Czech connective *přesto* "yet" has two equal German equivalents (*trotzdem, dennoch*).

Our conclusions support the assumption that there is usually no pure one-to-one correspondence between connectives across languages, which makes difficult the efforts of NLP processing of discourse relations expressed by connectives across languages. This finding can also be considered a further proof of the fact that some connectives have a broader meaning or broader possibilities of use in texts than other connectives with similar meaning (as presented in [16]).

5 CONCLUSIONS

In our paper, we compared using of discourse connectives by native and non-native speakers. We worked with the group of grammaticalized anaphoric connectives in Czech and German (originally coming from a combination of a preposition and an anaphoric expression).

Regarding the use of these connectives in texts by language learners, our material has not univocally shown that the frequencies of individual connectives linearly increase with the improvement of the author's language proficiency (in case of German, the selected connectives increase rather in jumps between the basic and independent language users). At the same time, the frequency of these connectives in the texts produced by non-native speakers is rather low, which demonstrates that language learners reach text coherence rather through other (probably not so complex) language devices.

It also turned out (both on Czech and German material) that the most frequent connectives are usually the same expressions in texts written by the native as well as non-native speakers. We can thus conclude that language learners acquire the strategies in using discourse connectives in accordance with the habits of native speakers. If a connective is widely used by native speakers, we may assume that it is also widely used by language learners.

The comparison of Czech and German language material demonstrated that the group of grammaticalized anaphoric connectives originally containing a preposition and an anaphoric expression is different in Czech and German. We detected about 15 connectives of this type in Czech and about 50 in German, i.e. German tends to create grammaticalized anaphoric connectives much more than Czech. This is supported also by the fact that the most frequent counterparts of some German one-word anaphoric connectives (like *dagegen* or *hierzu*) are Czech multiword anaphoric phrases (like *proti tomu* and *naproti tomu* or *za tím(to) účelem*, *v této souvislosti* and *vzhledem k tomu*).

The comparison of the pairs of Czech and German connectives showed that there is usually no pure one-to-one correspondence between connectives and their counterparts in other languages. The degree of equivalency of the individual connectives is highly variable: some pairs of expressions occur as equivalents in 90% of cases (cf. Czech *zatímco* and German *während*, both meaning “while”), whereas some only in 17% (cf. Czech *nadto* and German *überdies*, both meaning “moreover”). At the same time, the pairs of connectives are often influenced by the direction of translation (e.g. Czech *mimoto* is mostly translated as German *außerdem* but German *außerdem* is mostly translated as Czech *kromě toho*; all meaning “besides”). It means that some connectives have wider possibilities of use than the others (e.g. German *außerdem* covers the use of both Czech *mimoto* and *kromě toho*) and that connective equivalents in various languages are mostly only partial synonyms.

ACKNOWLEDGMENTS

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

REFERENCES

[1] A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová, C. Vettori, “The MERLIN corpus: Learner Language and the CEFR”. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik/Iceland: European Language Resources Association (ELRA), 2014.

[2] *Common European Framework of Reference for Languages (CEFR)*. Council of Europe, 2011. <https://www.coe.int/en/web/common-european-framework-reference-languages/>.

[3] V. Dovalil, T. Káňa, H. Peloušková, Š. Zbytovský, M. Vavřín, *Corpus InterCorp – German*. Version 10. Prague/Czechia: CNK FF UK, 2017. Cit. 4. 5. 2018. <http://www.korpus.cz>.

[4] *DWDS-Kernkorpus des 20. Jahrhunderts*. Version 3. Cit. 4. 5. 2018. <https://www.dwds.de/d/k-referenz#kern>.

[5] J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková-Razímová, Z. Urešová, *Prague Dependency Treebank 2.0*. Philadelhpia/USA: ÚFAL MFF UK, 2006. <http://www.ldc.upenn.edu>.

[6] M. Křen, V. Cvrček, T. Čapka, A. Čermáková, M. Hnátková, L. Chlumská, T. Jelínek, D. Kováříková, V. Petkevič, P. Procházka, H. Skoumalová, M. Škrabal, P. Truneček, P. Vondřička, A. Zasina, *Korpus SYN*, version 6. Cit. 4. 5. 2018. Praha/Czechia: FF UK, 2017. <http://korpus.cz>.

[7] A. Nedoluzhko, E. Lapshinova-Koltunski. “Correlating DRDs with other types of discourse phenomena. Cross-linguistic analysis of the interplay between DRDs, coreference and bridging”. *Cross-Linguistic Discourse Annotation. Applications & Perspectives*. Toulouse/France: Textlink, pp. 83–88, 2018.

[8] R. Pasch, U. Brauße, E. Breindl, U. H. Waßner, *Handbuch der deutschen Konnektoren*. Berlin/New York: Walter de Gruyter, 2003.

[9] L. Poláková, P. Jínová, J. Mírovský, “Interplay of Coreference and Discourse Relations: Discourse Connectives with a Referential Component”. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Istanbul/Turkey: European Language Resources Association (ELRA), pp. 146–153, 2012.

[10] L. Poláková, P. Jínová, Š. Zikánová, E. Hajičová, J. Mírovský, A. Nedoluzhko, M. Rysová, V. Pavlíková, J. Zdeňková, J. Pergler, R. Ocelák, *Prague Discourse Treebank 1.0*. Prague/Czechia: ÚFAL MFF UK, 2012. <https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0008-E130-A>.

[11] L. Poláková, P. Jínová, Š. Zikánová, Z. Bedřichová, J. Mírovský, M. Rysová, J. Zdeňková, V. Pavlíková, E. Hajičová, *Manual for Annotation od Discourse Relations in Prague Dependency Treebank*. TR 2012/47. Prague/Czechia: ÚFAL MFF UK, 2012.

[12] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber. “The Penn Discourse Treebank 2.0”. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech/Morocco: European Language Resources Association (ELRA), 2008.

[13] A. Rosen, M. Vavřín, A. J. Zasina, C*orpus InterCorp – Czech*. Version 10. Prague/Czechia: CNK FF UK, 2017. Cit. 4. 5. 2018. <http://www.korpus.cz>.

[14] M. Rysová, J. Mírovský, “Use of Coreference in Automatic Searching for Multiword Discourse Markers in the Prague Dependency Treebank”. *Proceedings of the 8th Linguistic Annotation Workshop (LAW-VIII)*. Dublin/Ireland: Dublin City University, pp. 11–19, 2014.

[15] M. Rysová, K. Rysová, “The Centre and Periphery of Discourse Connectives”. *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC)*. Bangkok/Thailand: Chulalongkorn University, pp.452–459, 2014.

[16] M. Rysová, K. Rysová, “Primary and Secondary Discourse Connectives: Constraints and Preferences”. *Journal of Pragmatics 130*, pp. 16–32, 2018.

[17] M. Rysová, K. Rysová, J. Mírovský, M. Novák, “Introducing EVALD – Software Applications for Automatic Evaluation of Discourse in Czech”. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*. Shoumen/Bulgaria, pp. 634–641, 2017.

[18] M. Rysová, P. Synková, J. Mírovský, E. Hajičová, A. Nedoluzhko, R. Ocelák, J. Pergler, L. Poláková, V. Pavlíková, J. Zdeňková, Š. Zikánová, *Prague Discourse Treebank 2.0*. Prague/Czechia: ÚFAL MFF UK, 2016. <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1905>.

[19] M. Stede, DiMLex: “A lexical approach to discourse markers”. *Exploring the Lexicon – Theory and Computation*. Alessandria: Edizioni dell’Orso, 2002.

[20] M. Stede, Y. Grishina, “Anaphoricity in Connectives: A Case Study on German”. *Proceedings of the Workshop on Cereference Resolution Beyond OntoNotex (CORBON)*, collocated with NAACL 2016. San Diego/California: Association for Computational Linguistics, pp. 41–46, 2016.

[21] M. Stede, A. Neumann, “Potsdam Commentary Corpus 2.0: annotation for discourse research”. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik/Iceland: European Language Resources Association (ELRA), 2014.

[22] M. Stede, C. Umbach, “Dimlex: A lexicon of discourse markers for text generation and understanding”. *Proceedings of the 17th international conference on Computational linguistics*. Volume 2, pp.1238–1242, 1998.

[23] M. Vavřín, *Treq*. Praha/Czechia: FF UK, 2015. <http://treq.korpus.cz>.

[24] B. Webber, M. Stone, A. Joshi, A. Knott, “Anaphora and discourse structure”. *Computational Linguistics 29 (4)*, pp. 545–587, 2003.

[25] *Wörterbuch der Konnektoren*. Mannheim/Germany: IDS. https://grammis.ids-mannheim.de/