OPEN ACCESS

Parallel corpus in analysing Czech spoken expressions and their equivalents in English, French, and Polish

Adrian Jan Zasina (Charles University, Prague)

ABSTRACT

This paper uses corpus data to analyse spoken expressions and discourse markers in Czech, applying these findings to corpus-based exercises for learners of Czech as a foreign language. The analytical section highlights the usefulness of parallel corpus in identifying suitable translation equivalents for prevalent Czech spoken vocabulary in English, French, and Polish as native languages from the learner's perspective. The methodology outlines the process of finding appropriate translation equivalents in film subtitles, considering both meaning and spoken register. The pedagogical section introduces three corpus-based exercises designed to improve conversational skills, featuring authentic texts that familiarise learners with spoken vocabulary. This research builds on previous studies of the English language that did not use parallel corpora to identify translation equivalents in learners' native languages — an essential factor for understanding a foreign language. In addition, tailor-made corpus-based exercises can be seamlessly integrated into everyday classroom activities to enhance language awareness among non-native speakers.

KEYWORDS

corpus, corpus-based exercises, Czech, data-driven learning, discourse markers, speaking skills, spoken expressions

DOI

https://doi.org/10.14712/18059635.2024.2.2

1 INTRODUCTION

The process of acquiring a foreign language requires substantial exposure to natural language to yield discernible results. Therefore, employing authentic linguistic material is crucial for facilitating the development of a learner's language skills, particularly for advanced learners aiming to attain fluency akin to that of native speakers, especially in conversational contexts. However, dialogues found in coursebooks often diverge significantly from real-life language use (Gilmore, 2004). Despite attempts to incorporate features typical of spoken language registers, these dialogues are artificial constructs created by coursebook authors (Holá, 2019). Consequently, they lack the authenticity inherent in everyday interactions, which include response tokens, natural repetitions, false starts, hesitations, etc. (Walsh, 2010). McCarthy and Carter (2001, p. 338) aptly emphasize the importance of appropriate input to achieve desirable language output: "Whatever else may be the result of imaginative methodologies for eliciting spoken language in the L2 classroom, there can be little hope for natural spoken output on the part of language learners if the

input is stubbornly rooted in models that owe their origin and shape to the written language."

Fortunately, today we have access to language corpora that provide examples of authentic language usage. Corpora have been extensively employed in language teaching for over 30 years. The approach of using corpus methods in learning was popularized by Johns (Johns, 1991) in the early 1990s. He coined the term *data-driven learning* (DDL) in his seminal work, describing this innovative method. According to Johns, students can discover language rules based on corpus evidence, thereby becoming independent in their learning. In recent years, numerous studies have employed the DDL method; however, the majority of these do not focus on developing speaking skills (Boulton & Cobb, 2017, p. 379). They primarily concentrate on writing, vocabulary, lexicogrammar, and grammar. Nevertheless, studies investigating typical spoken vocabulary, such as linking adverbials (Boulton, 2009), epistemic stance markers (Şahin Kızıl & Savran, 2018) or the discourse marker *well* (Huang, 2019), have emerged. Unfortunately, in the context of teaching Czech as a foreign language, such corpus studies are currently unavailable.

The sociolinguistic situation of Czech is sometimes considered as near-diglossic (Bermel, 2014), characterized by two distinct varieties: Literary Czech (*spisovná čeština*) and Common Czech (*obecná čeština*). Literary Czech represents the prestigious variant used in the majority of situations, while Common Czech is employed in everyday informal spoken communication. There is ongoing debate regarding the incorporation of Common Czech into foreign language teaching, with scholars holding differing perspectives on this matter. Some advocate for the introduction of spoken language elements even at the basic level (Holá, 2019, pp. 115–116), while others suggest its integration at an advanced level of communication competence (Hrdlička, 2019, p. 101). Undoubtedly, advanced learners should recognize the marked differences between spoken and written styles, especially when striving for a more "native" sounding expression. Nonetheless, working with authentic language samples is crucial for advanced learners to attain appropriate fluency in spoken interactions with native speakers.

Hence, this study aims to analyse spoken expressions that can later be used in teaching speaking skills in Czech as a foreign language. The primary goal is to conduct an analysis based on parallel corpus data to identify the most appropriate translation equivalents of frequently used Czech spoken vocabulary in English, French, and Polish, considering the learner's perspective as native speakers of these languages. To effectively apply the features of spoken language, learners need to first understand their meaning in their mother tongue. Barlow (2000, p. 110) notes that "learning a second language involves some use of first language schemas as templates for creation of schemas for the second language, as well as the formation of completely new schemas for the second language." Moreover, it is crucial to make learners aware of how to work with corpus data to obtain relevant translation equivalents. For this purpose, a subtitle component of a parallel corpus, emulating a spoken register, is used.

The secondary goal is to apply the vocabulary under examination to corpus-based exercises. These exercises aim to practice and reinforce spoken vocabulary while producing dialogues that enhance conversational skills. They consist of authentic lan-

guage examples derived from parallel, written, and spoken corpora, intended for use in classroom activities as hands-on exercises.

The paper is further divided into four sections. Section 2 presents data and methodology, describing the process of choosing the lexemes under examination. Section 3, devoted to corpus analysis, is the core part of the paper. Firstly, it comments on the semantics of spoken expressions and discourse markers, and the importance of using parallel corpus data in this study. Secondly, it provides an in-depth analysis of the subtitle corpus with regard to translation equivalents in English, French, and Polish, and underscores the importance of understanding the Czech spoken register through their equivalents derived from a parallel corpus. Thirdly, it summarizes the most common tendencies in translation strategies and offers general comments on working with quasi-spoken parallel data. Section 4 delivers ideas for corpus-based classroom activities that could improve the conversational skills of advanced Czech learners. It also highlights the usefulness of the DDL method in language teaching and learning. Section 5 discusses and concludes the main findings.

2 DATA AND METHODOLOGY

First, it was necessary to find candidates for spoken words for further analysis of typical spoken vocabulary in the parallel corpus. To compile an appropriate list of spoken vocabulary, I used frequency lists of the top 3,000 written and spoken Czech lemmas set by Škrabal, Laubeová and Štěpánková (2022, pp. 32–92). The list is a ready-to-use set based on written and spoken corpora. The lemmas in this list were marked with the letters W (for written) and S (for spoken), followed by a number 1–3, indicating the first, second, and third thousand. I examined a wordlist with the code S1 — the most frequent lemmas marked as spoken within the first thousand — which consists of 180 lexemes. To reveal a suitable vocabulary for the learners, further categorisations were needed (see 2.2 Method). For the purpose of analysis, a parallel corpus was used; for preparing corpus-based exercises, corpora of contemporary spoken and written Czech were used (see 2.1 Data).

2.1 DATA

This study makes use of data from the InterCorp release 12 parallel corpus (Čermák & Rosen, 2012) to establish a list of translation equivalents for typical spoken vocabulary. InterCorp consists of original texts and their translations in two main components, called the core and the collection. The core primarily contains fiction and is semi-manually aligned[1], whereas the collection is automatically annotated and is further divided into six components:

1 First, the core undergoes automatic alignment using the hunalign software (Varga et al., 2007; http://mokk.bme.hu/en/resources/hunalign/). Second, the automatically aligned texts are uploaded to the parallel editor InterText (Vondřička, 2014; https://wanthalf.saga.cz/intertext) where they are checked semi-manually.

— legal texts of the European Union from Acquis Communautaire corpus (https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis),
— translation of the Bible,
— political commentaries from Project Syndicate (https://www.project-syndicate.org/),
— political commentaries from VoxEurop (https://voxeurop.eu/),
— proceedings of the EU parliament from 2007–2011 (http://www.statmt.org/europarl/),
— film subtitles provided by the Open Subtitles database (https://www.opensubtitles.org/).

For the purpose of my study, I work with the subtitle component and its Czech (Rosen, Vavřín, & Zasina, 2019a), English (Klégr et al., 2019), French (Nádvorníková & Vavřín, 2019), and Polish (Bańczyk, Dybalska, & Vavřín, 2019) parts. The film subtitles were chosen because of their similarities to the spoken register. It is not feasible to obtain parallel texts of spoken language; however, subtitles can effectively emulate natural and spontaneous spoken communication. Therefore, in terms of parallel data, they are the most suitable material for analysing the most frequent spoken vocabulary across languages. Moreover, this approach makes it possible to identify translation equivalents in the languages examined. Of course, I am aware that this is not real spoken language (in this context, it is rather quasi-spoken language), so it is necessary to engage spoken corpora in the preparation of corpus exercises.

In terms of corpus-based exercises, the ORTOFON corpus (Kopřivová, Komrsková, Lukeš, Poukarová, & Škarpová, 2017) and SYN2015 corpus (Křen et al., 2015) were used. ORTOFON is a Czech spoken corpus of spontaneous (informal) everyday communication that is lemmatized, morphologically tagged, and balanced with respect to several sociolinguistic categories (Komrsková, Kopřivová, Lukeš, Poukarová, & Goláňová, 2017). It contains recordings of spontaneous conversations between people who have familiar or friendly relations; none of the conversations is experimentally induced. On the other hand, SYN2015 is a representative corpus of contemporary written Czech, consisting of lemmatisation, morphological, and syntactic annotation (Křen et al., 2016). It is balanced in terms of various text genres, including fiction, non-fiction, newspapers and magazines. Predominantly, it consists of texts published in the last five-year period (2010–2014). Both corpora are appropriate materials for providing suitable examples of spoken and written language to highlight their differences. Thus, this material is used in the corpus-based exercises presented in Section 4.

2.2 METHOD

The current study concentrates on spoken expressions intended for teaching Czech learners. Hence, it was imperative to select appropriate words reflecting the natural speech of native speakers. I opted to pinpoint typical spoken expressions[2] and

2 By spoken expressions, I mean lexemes that are frequent in the spoken register and have written counterparts with which they form pairs, such as: *cop — policeman*, *bro — brother*, *granny — grandmother*.

discourse markers (DM)[3] from a wordlist comprising 180 lexemes obtained from Škrabal, Laubeová and Štěpánková's frequency list (2022, pp. 32–92). Several identified categories do not meet the established criteria:

- exclamations (e.g., *ježišmarjá* 'Jesus Christ', *ježíš* 'geez', *jé* 'oh', *ahoj* 'hello', *fuj* 'yuck'),
- swear words (e.g., *prdel* 'ass', *hovno* 'shit', *kráva* 'bitch', *píča* 'pussy', *debil* 'asshole'),
- deixis (e.g., *tenhleten* 'this', *tadyhle* 'here', *tamhle* 'there', *nikam* 'nowhere'),
- hesitation (*hmm*, *aha*, *emm*),
- dialectal words (e. g. *aj* 'and', *tož* 'well'),
- words with a thematic focus (e.g., nouns: *záchod* 'toilet', *angličtina* 'English language', *chata* 'cottage', *doktor* 'doctor'; verbs: *sníst* 'to eat', *dojet* 'to reach', *vymyslet* 'to think up', *lézt* 'to climb'; adjectives: *strašný* 'terrible', *zvědavý* 'curious', *šílený* 'crazy', *napsaný* 'written'),
- other frequent parts of speech (e.g., numerals: *šedesát* 'sixty', *osmdesát*, 'eighty', *devadesát* 'ninety' adverbs: *akorát* 'only, just', *trošku* 'a little', *super* 'great', *hezky* 'nice'; conjunctions: *čili* 'so', *poněvadž* 'because').

Two more categories were identified: spoken expressions (Table 1) and DM (Table 2), which are used in the analysis presented in the following paragraph. Lexemes in the tables are ranked based on their frequency in spoken corpora. The tables also include translations into English, ipm (instances per million) in spoken corpora, ipm in the written corpus SYN2015, and the spoken/written (S/W) ratio of ipm. When comparing both ipm values, it is noticeable that these words are predominantly used in spoken language.

In terms of DM in Table 2, it must be highlighted that lexemes such as *myslet*, *normálně*, *vyloženě*, *kdyžtak*, *schválně*, *víceméně* have grammatical and semantic functions. However, they can also be considered DM in certain contexts, e.g., *Vypadá úplně normálně* (She looks totally normal) vs. *Normálně žárlíš* (You're simply being jealous), as commented on in the analytical part. It was not possible to differentiate between these two functions automatically; therefore, the numbers presented in the table are shown for the entire lemma, and the English translations correspond to the basic meaning.

In the analytical part, the Treq tool (Škrabal & Vavřín, 2017; Vavřín & Rosen, 2015) was employed to generate a list of the most frequent translation equivalents for Czech spoken expressions (Tables 4, 6, 7) and DM (Table 8) under examination. A maximum of three of the most frequent examples were chosen, except in one in-

3 DM are defined as "words or phrases that appear to have no grammatical or semantic function, such as *you know*, *like*, *oh*, *well*, *I mean*, *actually*, *basically*, *OK* as well as connectives like *because*, *so*, *and*, *but* and *or*" (Baker & Ellece, 2011, p. 34). Schiffrin (1987, p. 31) defines DM in her seminal work as "sequentially dependent elements that bracket units of talk". In this study, DM are understood as words with no grammatical or semantic function but with a pragmatic function; they are typical of spoken communication and are an important factor in maintaining a conversation. DM are involved in spontaneous interactions between interlocutors.

No.	Lemma	Translation	Frequency	ipm	SYN2015 ipm	S/W ratio
1	furt	still	6948	914.42	8.86	103.21
2	barák	house	1998	262.96	18.76	14.02
3	mamka	mum	1768	232.69	10.91	21.33
4	děcko	kid	1214	159.77	14.15	11.29
5	taťka	dad	1055	138.85	7.16	19.39
6	kafe	coffee	1046	137.66	17.05	8.07
7	sranda	fun	915	120.42	11.11	10.84
8	kilo	kilo	809	106.47	20.79	5.12
9	brácha	bro	764	100.55	14.46	6.95
10	ségra	sis	681	89.63	3.06	29.29
11	babi	granny	647	85.15	3.79	22.47
12	prachy	money	629	82.78	12.20	6.79
13	ženská	woman	563	74.10	28.73	2.58
14	chudák	poor	559	73.57	19.01	3.87
15	polívka	soup	417	54.88	6.37	8.62
16	policajt	cop	415	54.62	15.60	3.50
17	krám	shop	407	53.57	9.03	5.93
18	kámoš	mate	396	52.12	11.16	4.67
19	flaška	bottle	392	51.59	4.22	12.23
20	strejda	uncle	386	50.80	7.91	6.42
21	borec	guy	373	49.09	9.46	5.19
22	gympl	high school	371	48.83	2.51	19.45
23	obývák	living room	360	47.38	17.29	2.74
24	foťák	camera	347	45.67	5.07	9.01
25	baterka	torch	340	44.75	11.20	4.00
26	mail	email	338	44.48	9.72	4.58

TABLE 1. Spoken expressions

No.	Lemma	Translation	Frequency	ipm	SYN2015 ipm	S/W ratio
1	myslet	to think	16030	2109.71	680.07	3.10
2	viď	right	7353	967.73	14.47	66.88
3	hele	hey, look	6694	881.00	21.91	40.21
4	normálně	normally	5841	768.73	27.71	27.74
5	jakože	like	3310	435.63	2.08	209.44
6	jó	yeah	1453	191.23	2.72	70.30
7	cože	what	977	128.58	20.46	1285.83
8	hej	hey	956	125.82	8.49	14.82
9	vyloženě	strictly, downright	710	93.44	10.43	8.96
10	kdyžtak	just	679	89.36	0.31	288.27
11	holt	just, well	463	60.94	6.18	9.86
12	schválně	on purpose	349	45.94	10.64	4.32
13	víceméně	more or less	344	45.27	19.73	2.29

TABLE 2. Discourse markers

stance where four were selected. Translation equivalents were arranged based on absolute frequency for each language separately, and particular equivalents on the same line may not necessarily correspond to each other.

Example sentences from the InterCorp corpus were provided for the chosen lexemes to illustrate the most suitable spoken equivalents in all four languages. These sentences can be utilized in a classroom setting to demonstrate appropriate strategies used in spoken communication.

In some cases, on-line dictionaries were consulted to verify whether the translation equivalents were identified as informal or spoken words. The English *Cambridge Dictionary* (https://dictionary.cambridge.org/), the French *Le Dictionnaire* (https://www.le-dictionnaire.com/), and the Polish *Słownik języka polskiego PWN* (https://sjp.pwn.pl/) were used for this purpose.

3 ANALYSIS

This paragraph presents an analysis conducted on data from InterCorp. The objective of the examination is to identify the most suitable translation equivalents of spoken Czech expressions and DM (Tables 1 and 2) in English, French, and Polish, considering them from the perspective of the learner's native languages. The aim is not only to find any equivalent but also to locate appropriate translation equivalents that suit the spoken register. The results should assist teachers in approaching corpus data and understanding the types of issues that must be considered when working with corpus evidence. Moreover, the obtained translation equivalents may be directly applied to classroom activities, and examples derived from the analysis could provide learners with authentic usage illustrations.

This paragraph is divided into three subsections. Subsection 3.1 discusses the examined lexemes in terms of their meaning and function. Subsection 3.2 focuses on the subtitle material, explaining how to interpret corpus instances and analysing the obtained results. Subsection 3.3 recapitulates the main findings and formulates general tendencies in the investigated languages.

3.1 SPOKEN EXPRESSIONS AND DISCOURSE MARKERS

Looking at the category of spoken expressions, it is possible to identify designations of family members (e.g., *mamka*, *děcko*, *taťka*), people (e.g., *ženská*, *chudák*, *policajt*), buildings (e.g., *barák*, *krám*, *gympl*), food and drinks (*kafe*, *polívka*), items (*flaška*, *foták*, *baterka*), and other colloquial expressions (*furt*, *sranda*). This primarily highlights the thematic focus of typical conversation. Within the DM category, it is possible to identify verbs of thinking (*myslet*), confirmation checks (*viď*, *cože*), phatic expressions (*hele*, *jó*, *hej*), fillers (*jakože*, *kdyžtak*), boosters (*vyloženě*), downtoners (*víceméně*).[4] They are important factors to maintain a conversation.

4 The division of DM was inspired by the lists of language futures (Cvrček, Laubeová, et al., 2020b) and the chapter on DM (Čermáková, Jílková, Komrsková, Kopřivová, & Poukarová, 2019).

These types of expressions are not typically included in coursebooks; however, learners may frequently encounter them in interactions with native speakers.[5] Foreigners might face challenges in comprehending and actively using such spoken expressions, as well as in finding suitable translations that correspond to spoken communication in their native languages (in this case, English, French, and Polish). Furthermore, printed dictionaries often fail to provide translation equivalents for all potential situations. They struggle to keep pace with evolving language use and cannot encompass equivalences that have emerged in recent years. Fortunately, modern learners now have the opportunity to acquire vocabulary through online access to corpus data.

The collection of subtitles emulates spoken language that cannot typically be obtained in parallel texts. The analysis does not focus on the direction of translation but rather on the examples themselves. The direction of translations is not considered here because, in this study, capturing appropriate equivalents typical of the spoken register is more important than addressing interference from the source language. However, it is worth noting that most film subtitles are translated from English into other languages.

3.2 SUBTITLE SUBCORPUS

It is important to carefully select examples from the subtitle subcorpus, as it contains a large number of irrelevant instances that do not provide a meaningful situational context for non-native speakers. Sometimes, the concordance displays sentences that are too short or where the translation is misaligned (see Figure 1). In such cases, it is necessary to click on the sentence to expand the context.

_SUBTITLES	- Ahoj , **mamko** .	_SUBTITLES	- Hey , Mama .
_SUBTITLES	**Mamka** …	_SUBTITLES	Ah , do n't worry about it .
_SUBTITLES	**Mamka** ji používá .	_SUBTITLES	She , uh … She usin ' it .
_SUBTITLES	- Jo , **mamko** .	_SUBTITLES	- " Yeah , Mom . "
_SUBTITLES	Tady **mamka** .	_SUBTITLES	It 's Mommy .

FIGURE 1. Parallel concordance of lexeme *mamka* 'mum' in the subtitle collection

Teachers should also be aware that the translations from the Open Subtitles database are not authorized by any official film company. Therefore, it is crucial for teachers to choose examples wisely for their classroom activities. It is recommended to prepare material in advance to avoid any misunderstandings, especially when working with learners at a lower proficiency level, which requires careful preparation. However, the subtitles represent a vast collection of data that allows us to identify the most frequent occurrences. This frequency-based approach helps avoid unnecessary instances. Consequently, the following analysis deals with the most frequent trans-

5 As Bulejčíková (2015, p. 76–86) concludes, Czechs tend to use quite often Common Czech in communication with foreigners (in more than a quarter of the records analysed).

lation equivalents and outlines steps for approaching corpus data and interpreting results from the subtitle collection. To highlight the most appropriate strategies for learners engaged in spoken communication, it is essential to examine the chosen corpus instances, focusing on the most suitable spoken equivalents.

3.2.1 SPOKEN EXPRESSIONS

The first group of spoken expressions under analysis comprises a list of family members: *mamka*, *děcko*, *taťka*, *brácha*, *ségra*, *babi*, *strejda*. It is important to be aware that each language pair (e.g., Czech-English or Czech-French) in InterCorp contains a different number of texts. Therefore, the number of instances varies for each lexeme depending on the language component. The table below shows the absolute frequency of these family members in the subtitle collection for the Czech-English, Czech-French, and Czech-Polish components.

Lexeme / **Component**	***mamka***	***děcko***	***taťka***	***brácha***	***ségra***	***babi***	***strejda***
Czech-English	1,111	3,029	1,008	4,717	554	1,001	1,373
Czech-French	592	1,634	540	2,425	277	420	650
Czech-Polish	726	2,136	653	3,421	365	625	915

TABLE 3. Number of hits for family members in the subtitle collection of InterCorp

No.	**Family member**	**English Translation**	**French Translation**	**Polish Translation**
1.	mamka	mom	mère	mama
		mother	maman	mamusia
		mum		matka
2.	děcko	kid	enfant	dziecko
		baby	bébé	dzieciak
		child	gosse	
3.	taťka	daddy	papa	tata
		dad	père	tatuś
		papa		tato
4.	brácha	brother	frère	brat
		bro	frangin	stary
		man	pote	brach
5.	ségra	sister	sœur	siostra
		sis	sœurette	siostrzyczka
			frangine	
6.	babi	grandma	mamie	babcia
		granny	mémé	babunia
7.	strejda	uncle	oncle	wujek
			tonton	wuj

TABLE 4. Translation equivalents for the family members group, based on the Treq tool

Based on Table 3, it can be assumed that the number of translation equivalents will vary and may not be consistent across the three chosen languages. However, common tendencies are observed in the frequency of lexeme usage for each language, with the most frequent being *brácha*, *děcko*, and *strejda*. Table 4 presents a list of the most frequent translation equivalents for each lexeme.

It is noticeable that some lexemes have both formal and informal translation equivalents (e.g., *děcko*, *brácha*, *ségra*, *strejda*, and for French also *mamka*, *taťka*), while others are only informal for English and French (e.g., *babi*) and only formal for English (e.g., *strejda*). This highlights differences in language strategies; for instance, there is no typical informal equivalent in English (based on the corpus evidence) for *strejda*. It also underscores that Polish and Czech have the most in common, as Polish does not have formal equivalents for *mother* as English and French do.

(1) CS Mám počkat na **mamku**.
EN I'm supposed to wait for **mom**.
FR Je dois attendre **maman**.
PL Powinienem poczekać na **mamę**.

To illustrate spoken translation equivalents for *mamka* in all four languages, Example (1) is provided. Based on the Treq results, the most applicable equivalents are: *mom* in American English (or *mum* in British English), *maman* in French, and *mama* in Polish. All these equivalents represent a similar register and are stylistically comparable. The same applies to lexemes such as *taťka* and *děcko*, with their spoken equivalents being *dad* and *kid* in English, *papa* and *gosse* in French, and *tata* and *dzieciak* in Polish.

(2) CS Chodí na vejšku s mým **bráchou**.
EN He goes to college with my **brother**.
FR Il est à la fac avec mon **frère**.
PL Studiuje z moim **bratem**.

(3) CS Alane, těší mě, **brácho**.
EN Alan, nice to meet you, **bro**.
FR Alan, ravi de te faire ta connaissance, mon **pote**.
PL Alan, miło cię poznać, **stary**.

A more complex example is the lexeme *brácha*, which can have two distinct meanings: (2) a male person who shares the same parents, and (3) a male friend. For the first meaning, the formal equivalents *brother*, *frère*, *brat*, along with their familiar counterparts *bro* and *fragin*, are most commonly used based on the Treq results. For the second meaning, the familiar counterparts are interchangeable with spoken expressions such as *man* (or *mate* in British English), *pote*, *stary*, and *brach* (used only in the vocative case). This meaning is related to the lexeme *kámoš*, which will be discussed later.

For the lexeme *ségra*, the data show that formal equivalents are more frequently used, such as *sister*, *sœur*, and *siostra*, while familiar expressions like *sis*, *sœurette* are used less often. In Polish, there were a few rare instances of *siora* that could also be appropriate here in this context, although it was seldom used.

(4) CS Běž počkat do auta, **babi**.
EN Go wait in the car, **grandma**.
FR Va dans la voiture, **mamie**.
PL Poczekaj w samochodzie, **babciu**.

In the case of the lexemes *babi* (4) and *strejda*, Polish uses more frequent familiar terms (*babcia*, *wujek*), while English and French prefer different approaches. For *babi*, English and French use familiar equivalents, whereas for *strejda*, English tends to use a formal equivalent, and French also predominantly opts for a formal term.

The next group of spoken expressions consists of names of people: *ženská*, *chudák*, *policajt*, *kámoš*, and *borec*. The table below shows the absolute frequency of these expressions in the subtitle collection for the Czech-English, Czech-French, and Czech-Polish components.

Lexeme / **Component**	***ženská***	***chudák***	***policajt***	***kámoš***	***borec***
Czech-English	4,564	2,265	3,964	15,218	563
Czech-French	2,421	1,213	2,208	7,908	285
Czech-Polish	3,032	1,482	2,884	11,226	391

TABLE 5. Number of hits for the names of people in the subtitle collection of InterCorp

No.	People	English Translation	French Translation	Polish Translation
1.	ženská	woman	femme	kobieta
		girl	fille	dziewczyna
		lady		baba
2.	chudák	poor + a noun	pauvre + a noun	biedny + a noun
		loser		biedak
3.	policajt	cop	flic	glina
		police	police	policjant
		policeman	policier	gliniarz
4.	kámoš	man	pote	stary
		friend	ami	kumpel
		buddy	mec	przyjaciel
		mate	vieux	kolega
5.	borec	guy	type	koleś
		man	homme	facet
		boy	gars	gość

TABLE 6. Translation equivalents for names of people, based on the Treq tool

The most frequent lexemes are *kámoš*, *ženská*, *policajt*, indicating a thematic concentration in the subtitle collection. A list of the most frequent translation equivalents for each lexeme has been established (see Table 6).

Similarly to the case of family members, it is possible to observe a comparable tendency with these lexemes, where some have both formal and informal translation equivalents (*ženská*, *policajt*, *kámoš*), while others are only informal (*borec*). The translations for the lexeme *ženská* highlight that the same stylistic layer is not present in English, French, and Polish as it is in Czech. Thus, English, French, and Polish use neutral equivalents (e.g., *woman*, *girl*, *lady*, *femme*, *fille*, *kobieta*, *dziewczyna*) that do not convey the spoken expression's nuance. However, Polish does have an informal lexeme, *baba*, which is stylistically closer to Czech *ženská*. Example (5) shows the most frequent translation equivalents for *ženská*, where the standard form is *žena*.

(5) CS Já jsem tu **ženskou** vlastně nikdy nepotkala.
EN Well, I've never actually met the **woman**.
FR Je n'ai jamais rencontré cette **femme**.
PL Właściwie nigdy nie widziałam tej **kobiety**.

An interesting example is the lexeme *chudák*, for which English and Polish have one-word equivalents. However, the construction of the adjective *poor* with a noun is used more frequently across all three languages (see Example 6). It is also important to note that the English equivalent *loser* is used exclusively in a negative sense, whereas the Czech *chudák* and Polish *biedak* can also be used in contexts where one expresses sympathy or regret for someone else.

(6) CS **Chudák**.
EN Poor **kid**.
FR Pauvre **gosse**.
PL Biedny **dzieciak**.

The most frequent equivalents for the lexemes *policajt*, *kámoš*, and *borec* are appropriate for the spoken register and align with the style of informal communication. For *policajt* and *kámoš*, formal equivalents are also attested but are less frequent (e.g., *policeman*, *policier*, *policjant*, *friend*, *ami*, *przyjaciel*). All translation equivalents for *borec* represent informal usage. Example (7) presents the equivalents for the lexeme *policajt*, for which spoken translations have been found in English, French, and Polish.

(7) CS Jak mám vědět, že nejste **policajt**?
EN How do I know that you're not a **cop**?
FR Comment je sais que vous n'êtes pas **flic** ?
PL Skąd wiem, że nie jesteś **gliną**?

The final group of spoken expressions consists of lexemes that were not further categorized: *furt*, *barák*, *kafe*, *sranda*, *kilo*, *prachy*, *polívka*, *krám*, *flaška*, *gympl*, *obývák*, *foťák*,

baterka, and *mail*. The most frequent translation equivalents for each lexeme have been extracted (see Table 7).

No.	Other expressions	English Translation	French Translation	Polish Translation
1.	furt	still	toujours	ciągle
		always	encore	nadal
2.	barák	house	maison	dom
		building	baraque	budynek
		place	immeuble	barak
3.	kafe	coffee	café	kawa
		joe	kawa	kawka
4.	sranda	to kid	drôle	żartować
		fun	marrant	zabawny
		funny	amusant	jaja
5.	kilo	pound	kilo	kilo
		kilo	livres	kilogram
6.	prachy	money	l'argent	pieniądz
		cash	fric	kasa
		dough	pognon	forsa
7.	polívka	soup	soupe	zupa
8.	krám	store	boutique	sklep
		shop	magasin	rzecz
		thing	truc	grat
9.	flaška	bottle	bouteille	butelka
				flaszka
10.	gympl	high school	lycée	liceum
				ogólniak
11.	obývák	living room	salon	salon
12.	foťák	camera	appareil (photo)	aparat
			caméra	kamera
13.	baterka	flashlight	lampe de poche	latarka
		battery	batterie	bateria
		torch	(lampe) torche	akumulator
14.	mail	mail	mail	mail
		email	email	email
			courriel	

TABLE 7. Translation equivalents for other spoken expressions, based on the Treq tool

The results in Table 7 demonstrate that the lexemes *barák*, *kilo*, and *prachy* have both formal and informal translation equivalents. *Barák*, in addition to its formal translations *house*, *maison*, and *dom* (see Example 8), is also translated into French and Polish as *barrack*. The lexeme *kilo* has the same spoken equivalent across all languages, with a formal variant that highlights cultural differences: *pound* and *livres* for English

and French, and *kilogram* for Polish. For *prachy*, the formal equivalent *money* (*l'argent*, *pieniądz*) is the most frequent in all three languages. However, other translations of this lexeme reflect the spoken register (*cash*, *fric*, *kasa*).

(8) CS Líbí se mi váš **barák**.
EN I like your **house**.
FR J'aime votre **maison**.
PL Podoba mi się pana **dom**.

The next two lexemes, *flaška*, *gympl*, illustrate cases where all three languages have a standard equivalent, while Polish has an additional typical spoken variant. Unfortunately, the parallel corpus data do not provide any typical spoken expressions for *flaška*, *gympl* in English and French. This absence does not imply that such equivalents do not exist, but rather that they were not captured in the data. For the lexeme *kafe*, spoken expressions were noted but were rare. Thus, Example (9) presents the lexeme *kafe* with spoken equivalents in English and Polish: *joe* (typically used in American English) and *kawka*. The French example includes *jus* ('juice'), but the corpus evidence attests a more frequent spoken variant *kawa*.

(9) CS Á, to je jiný **kafe**.
EN Ah, that's good **joe**.
FR Ça, c'est du **jus**.
PL Dobra **kawka**.

No data were found on typical spoken equivalents for the lexemes *polívka* and *obývák*. This absence may be attributed to their characteristics specific to the Czech spoken register: the process of narrowing *é* into *í* (e.g., *polévka* → *polívka*) and univerbation (e.g., *obývací pokoj* → obývák).

The case of the lexemes *krám* and *baterka* is interesting because they have dual meanings, which affects their translations. *Krám* can denote a place where one can buy something (e.g., *shop*, *store*, *boutique*, *magasin*, *sklep*) or refer to an unneeded item of little value (with informal variants such as *truc*, *grat*, and neutral variants like *thing*, *rzecz*). In contrast, *baterka* primarily means a *battery* or a *torch* (a *flashlight* in American English).

The next example is the lexeme *furt*, which has standard equivalents in English and French, while in Polish, *ciągle* was found, which tends to be more of a spoken expression, as illustrated in Example (10).

(10) CS Jseš **furt** na vejšce?
EN Are you **still** at the university?
FR **Toujours** à l'université ?
PL **Ciągle** jesteś na uniwersytecie?

The lexeme *sranda* can be translated as an adjective into all examined languages (e.g., *funny*, *drôle*, *zabawny*). In English and Polish, it also has verbal (e.g., *to kid*, *żartować*)

and nominal counterparts (e.g., *fun*, *jaja*). In Polish, the nominal translation is actually the phrase *robić sobie jaja*, which corresponds to the Czech phrase *dělat si srandu*, as illustrated in Example (11).

(11) CS Děláš si **srandu**?
EN You **kidding**?
FR Quoi? T'es **drôle**.
PL **Jaja** sobie robisz?

The final two examples involve the lexemes *foťák* and *mail*, the standard Czech terms are *fotoaparát* and *e-mail*. The most frequent translation equivalents in French and Polish are informal one-word variants *appareil* and *aparat*, respectively, which are shortened forms of *appareil photo* and *aparat fotograficzny*. In the spoken register, the context often makes it clear that *appareil and aparat* refer to a *camera*. For English, the equivalent *camera* can denote either a photo or video camera. The French equivalent *caméra* and Polish *kamera* specifically refer to video cameras. Regarding *mail*, the most frequent equivalent across all examined languages is *mail*, with *email* as secondary option. In French, the domestic variant *curriel* is also used.

3.2.2 DISCOURSE MARKERS

Finally, a list was compiled of the most frequent translation equivalents for the category of DM represented by the following lexemes: *myslet*, *viď*, *hele*, *normálně*, *jakože*, *jó*, *cože*, *hej*, *vyloženě*, *kdyžtak*, *holt*, *schválně*, *víceméně* (Table 8).

Thinking verbs are significantly prevalent in spoken language, playing a crucial role in expressing opinions and inquiring about others' views. The translation equivalents of the verb *myslet* in English, French, and Polish also represent verbs of thinking. Example (12) demonstrates the most frequent equivalents: *to think* in English, *penser* in French, and *myśleć* in Polish. This verb often appears in its pragmatic function as a DM in various forms, such as *myslím* (or *myslim*, *myslíš*), indicating its role in conveying one's thoughts or seeking others' opinions.

(12) CS To je přirozeně to místo, kde si **myslím**, že bys mi měl přestat říkat detaily.
EN That's a natural place, I **think**, to stop giving me details.
FR Je **pense** que ce serait bien d'arrêter de me raconter tout.
PL To normalne, **myślę**, że wystarczy mi tych szczegółów.

Viď is a DM used by speakers to draw attention or to seek approval, whether verbal or non-verbal. In English and French, informal equivalents such as *right*, *huh* and *hein*, *non* are used. For Polish, while the most frequent equivalent *prawda* is a standard variant, the informal *co* aligns better with the spoken register. This is illustrated in Example (13), where *right*, *hein*, and *co* are shown to be the most appropriate translations to maintain the spoken quality of the original Czech expression.

No.	Discourse markers	English Translation	French Translation	Polish Translation
1.	myslet	to think	penser	myśleć
		to mean	croire	sądzić
		to guess	réfléchir	mieć na myśli
2.	viď	right	hein	prawda
		huh	non	co
			pas vrai	tak
3.	hele	to look	regarder	hej
		hey	écouter	(po)słuchać
		to listen	hé	patrzeć
4.	normálně	normally	normalement	normalnie
		normal	normal	zwykle
		usually	habitude	zazwyczaj
5.	jakože	like	comme	jakby
6.	jó	yeah	ouais	tak
				taa(a)
7.	cože	what	quoi	co
8.	hej	hey	hé	hej
		yo	hey	hey
		yeah	salut	cześć
9.	vyloženě	strictly	absolument	szczególnie
10.	kdyžtak	if	–	w razie czego
		just		jakby co
11.	holt	just	–	–
12.	schválně	on purpose	exprès	celowo
13.	víceméně	more or less	plus ou moins	mniej więcej

TABLE 8. Translation equivalents for DM, based on the Treq tool

(13) CS Seš dost slavnej, **viď**?
EN Hey, you're pretty famous, **right**?
FR T'es célèbre, toi, **hein** ?
PL Jesteś dość znany, **co**?

The lexemes *hele* and *hej* are primarily used to draw attention, often at the beginning of an utterance. *Hele* is commonly translated as an imperative in the three examined languages: *look*, *regarde*, *(po)słuchaj*. In contrast, *hej* is more accurately translated as *hey*, *hé*, or *hej*, which align better with the spoken register. Example (14) demonstrates the most suitable translation equivalents for the lexeme *hele*.

(14) CS **Hele**, něco ti řeknu.
EN **Look**, I've got some news for you.
FR **Regardez**, j'ai des nouvelles pour vous.
PL **Słuchaj**, mam dla ciebie wiadomość.

The lexeme *normálně* has three equivalents in each language, but the most common translations are *normally*, *normalement*, *normalnie*. These equivalents correspond closely with each other, as illustrated in Example (15).

(15) CS On by pro mě **normálně** nehnul ani prstem.
EN You know, and **normally** he wouldn't lift a finger to do anything for me.
FR **Normalement**, il ne lèverait pas le petit doigt pour moi.
PL **Normalnie** on nie kiwnąłby nawet palcem.

The lexemes *jakože*, *cože*, *vyloženě*, *schválně*, and *víceméně* each have only one translation equivalent in English, French, and Polish. While these expressions often do not add significant content to a conversation, they are characteristic of spoken language. They provide speakers with additional time to organize their thoughts and structure their utterances.

The lexeme *jó* is a particularly interesting case. It translates into familiar variants such as *yeah* in English, and *ouais* in French. However, in Polish, the most frequent equivalent is the standard *tak*, with *taa* being a more familiar variant that aligns with the spoken register.

The last two lexemes, *kdyžtak* and *holt,* are quite specific to Czech. While finding an equivalent for these lexemes in English (such as *just*) is relatively straightforward, it is more challenging to identify suitable translations in French and Polish. No direct translations were found for *kdyžtak* and *holt* in French. For Polish, however, a translation equivalent for *kdyžtak* has been identified, as shown in Example (16). These observations underscore the difficulty in finding precise counterparts for certain Czech discourse markers in other languages, particularly in the case of less directly translatable terms.

(16) CS **Kdyžtak** budu mluvit já.
PL **W razie czego**, ja się tym zajmę.

3.3 SUMMARY

The analysis of subtitle material has indicated a few relevant approaches to working with quasi-spoken parallel data. First, the corpus data have provided a relevant source of typical translation equivalents for spoken language. Thanks to the wide variety of subtitle material in Czech, English, French, and Polish, the most frequent equivalents reflect not only accurate translations but also a suitable register. This information is crucial for learners of Czech who want to improve their speaking skills and fluency.

Second, the results of the analysis have revealed that Polish, in comparison with English and French, more often opts for familiar variants. This shows that Polish and Czech have more in common not only because of their genetic similarities but also because of their similar tendency to use familiar terms rather than official ones.

Third, a user of parallel corpora should be aware that the highest frequency may not always indicate a translation equivalent suitable for a particular situation (in this

case, the spoken register). Sometimes it is necessary to analyse the context to decide if a given equivalent meets our requirements; for example, in the case of the lexeme *prachy*, more appropriate equivalents appeared in the second position.

Fourth, in some cases, it is not possible to find any translation equivalent. There could be two reasons for this. Sometimes it is a question of data availability; too little data cannot provide enough translation examples to identify any equivalent. In other cases, it is a matter of language resources, which may be lacking in the target language, as in the example of the lexeme *holt*, for which no single reference in French or Polish was found.

The analysis of subtitle collections is crucial in language learning and teaching. These results should be taken into account when considering the implementation of corpus methods in teaching conversational skills in the Czech language. It is feasible to build on the identified spoken expressions and DM to prepare suitable exercises that enrich learners' knowledge of spoken Czech. Furthermore, the analysis of translation equivalents provides examples in the learners' native language or a well-known second language, offering essential information for understanding the Czech spoken register. The issues covered in this research should be incorporated into activities that may help learners master typical spoken expressions. Thus, the following section focuses on the introduction of concrete corpus-based exercises.

4 CORPUS CLASSROOM EXERCISES[6]

This section presents ideas for corpus-based classroom activities designed to enhance conversational skills of advanced learners of Czech. The first two exercises centre around spoken expressions described in Table 1, while the third exercise focuses on discourse marker vocabulary presented in Table 2. All exercises are tailored for advanced learners at a proficiency level of C1 or higher.

The first task involves material from the subtitle subcorpus of InterCorp release 12 (Rosen, Vavřín, & Zasina, 2019b). The objective is to introduce new spoken vocabulary to learners, requiring them to replace underlined expressions with formal or neutral equivalents familiar from their prior education at lower levels. All sentences are drawn from the subtitle subcorpus, which emulates spoken Czech. Subsequently, to practise these new expressions, learners collaborate in pairs to prepare a dialogue. They then perform these roles in front of the class, with the teacher providing assistance and correcting any mispronunciations after the learners' presentation. This exercise serves to acquaint learners with the topic of spoken expressions.

6 The corpus-based exercises presented here were published in a workbook (Zasina, 2023, p. 86–87).

Exercise 1. Podívejte se na níže uvedené věty ze subkorpusu filmových titulků korpusu InterCorp v12. Nahraďte potržené mluvené varianty výrazů jejich formálními nebo neutrálními ekvivalenty nacházejícími se v rámečku. Následně s partnerem zkuste vytvořit dialog s použitím těchto mluvených výrazů.

pořád, dům, dítě, máma, kilogram, káva, legrace, táta, sestra, žena, bratr, polévka

1. Furt to ještě bolí. 2. V tak velkém baráku jsem ještě nebyl. 3. Má tu děcko, máme být potichu. 4. Mamka nemá ráda růžovou. 5. Zhubla jsem o 2 kila. 6. Možná se stavím večer na kafe. 7. Tohle už není žádná sranda! 8. Podívejte, co mi taťka koupil k promoci. 9. Jseš ta nejlepší ségra, kterou mám. 10. Ta ženská mi bude tolik chybět. 11. Tvůj brácha na tebe moc tlačí. 12. Přinesl jsem vám trochu čínský polívky.

The second task underscores the discernible distinctions between written and spoken language, with the objective of elucidating a clear contrast to foster learners' awareness of employing appropriate vocabulary in specific contexts. This exercise utilizes sentence examples representing typical spoken utterances from the ORTOFON v1 spoken corpus (Kopřivová et al., 2017) and typical written utterances from the SYN2015 written corpus (Křen et al., 2015). Consequently, each sentence serves as an authentic illustration of real natural language usage.

Learners engage with word pairs, requiring them to choose between spoken and neutral/formal expressions to fill gaps based on contextual cues and characteristic features of text style that denote a particular register. Through this task, learners have the opportunity to enhance their ability to recognize the distinctive features of spoken and written styles.

Exercise 2. Která varianta je vhodná pro psaný a která pro mluvený styl? Doplňte správný výraz podle charakteristických vlastností stylu textu.

legrace × sranda
Zítra si možná užiji .....
No vidíš, to je přesně vono, že si z něj dělaj lidi .....
Občas, ale spíš si ze mě lidé dělají ..... nebo pokřikují Oto.

peníze × prachy
Tam jde vo ..... a lidi... se teď nechtěj učit novejm věcem.
Část ..... získaných prodejem složila před odjezdem u faráře.
Stát nemusel vyplácet ....., které tak jako tak neměl.

kamarád × kámoš
Slyšelas to vod jednoho ....., že to říkal, a teď jsem to četla i v knížce.
Tři ..... ho stále tiše pozorovali.
Tak ten jeho ..... to je nějakej několikanásobnej železnej muž v Čechách.

fotoaparát × foťák
Budete-li si ..... vybírat v obchodě, nezapomeňte si vyzkoušet právě video.
Mě to hrozně nebavilo fotit... navíc mi přišlo že ten ..... je nějakej špatnej.
Ve výbavě nechybí 5Mpx ....., A-GPS nebo digitální kompas.

policisté × policajti
Když ..... nehodu vyšetřovali, někdo je nasměroval právě na domnělého pachatele.
Prostě zavolám ..... a nechám si to prostě zaplatit od silnic, že?
A tak se prej na sebe tak ti ..... podívali a teprv pak jakože rychle naskočili do auta.

The third exercise is centred on an authentic dialogue extracted from the ORTOFON corpus (Kopřivová et al., 2017), featuring a substantial use of DM. Initially, learners have the opportunity to immerse themselves in a genuine conversation and identify DM based on a provided list in the included box. The dialogue can be read independently by learners or read aloud in the classroom by selected students. Following the reading, learners are allotted time for text analysis.

Subsequently, the task involves contemplating the function of DM within the text. Learners collaborate in pairs to discuss their findings before engaging in an open class discussion to share individual insights. In the third phase, learners are tasked with applying the information acquired in practical terms. They work in pairs to prepare a dialogue, specifically utilizing DM not present in the provided example dialogue. Learners may seek assistance from the teacher or consult corpus data to explore additional instances of DM usage. Finally, they present their dialogues to their classmates.

Exercise 3. Přečtěte si pasáž neformální konverzace z korpusu ORTOFON a pokuste se na základě níže uvedené tabulky označit diskurzní markery v dialogu. Jakou plní roli v textu? Společně s partnerem připravte dialog s použitím dalších diskurzních markerů, jež se v dialogu neobjevily.

myslet, viď, hele, normálně, jakože, jó/jo, cože, hej,
vyloženě, kdyžtak, holt, schválně, víceméně

Martina: *(...) To se mi stalo o prázdninách, že jsem si otevřela, že jsem nějaké téma našla a zajímalo mě... Tak jsem to otevřela, Mladou frontu i Právo, a měli úplně kompletně stejný článek...*
Veronika: *Jako úplně doslova?*
Martina: *... jenom měli jinou fotku... hmm*
Veronika: *Ty jo!*
Martina: *Jakože no doslova jakože... některé ty pasáže byly fakt doslovné... některé byly jakože malinko jiné, ale jakože fakt se to lišilo víceméně v obrázku.*
Veronika: *Ty jo!*

Martina: *Takže to by bylo fakt jako docela drsné, protože já jsem četla nejprve Mladou frontu. A pak jsem si říkala, tak jako zkusím se ještě podívat třeba do tých jakože do Hospodářek a do Práva a… normálně to na mě vyjelo to samé.*
Veronika: *Jo, jo, jo, jo.*
Martina: *A teďka člověk řekne, ty jo, však jsem to před chvílí četla toto… jako hmm.*
Veronika: *Tak to je zvláštní jako.*
Martina: *Ale je to zajímavé dělat to o tom kraji.*
Veronika: *Já mám strašnou žízeň normálně… po tom… výstupu tam.*
Martina: *Já jsem myslela, že po víkendu.*
Veronika: *To ne, to měla Renča už včera.*
Martina: *To měla Renča.*
Veronika: *Jak ona furt pila a pila a já říkám, ty jo, to máš ještě z té soboty určitě a ona: myslíš?*
Martina: *Však já jsem jí to taky říkala, já jsem si tam dělala šťávu v kuchyni… a ona říkala já jsem měla strašnou žízeň, včera já jsem vypila během rána dva litry (…)*

The proposed exercises can be modified by each teacher to adapt them to the needs of learners. The exercises are also suitable as an introduction to the topic of spoken Czech or as revision material. The corpus evidence included in all three exercises provides clear information about the differences between spoken and written language.

5 CONCLUSION

Bearing in mind the complex situation of the Czech language bordering on diglossia, this research attempts to analyse spoken expressions and DM that are crucial in communication with native speakers and might later be used to teach speaking skills in Czech as a foreign language. The analysis indicates that the subtitle component of the parallel corpus may be a helpful source for examining translation equivalents of typical spoken vocabulary. The quasi-spoken parallel data enable the identification of relevant equivalents in English, French, and Polish that correspond to the spoken register as well. This is an especially valuable finding because it is not obvious to relate parallel data with spoken language, as they normally consist of fiction and non-fiction. The subtitle collection certainly brings benefits; however, it is necessary to take into account the potential bias that might influence the results. For instance, the most frequent translation equivalent may not always be appropriate for spoken language. Therefore, a linguist's expertise and context analysis are also required. Furthermore, corpus-based exercises demonstrate how to approach the teaching features of the spoken register using corpus data. All presented exercises are comprised of authentic texts representing written, quasi-spoken, and spoken Czech. They effectively introduce learners to the topic of spoken Czech and provide clear information regarding differences and vocabulary usage. Moreover, they compel learners to practise oral production and utilize spoken expressions and discourse markers to sound more natural when speaking. Particularly valuable is the

situational context delivered by corpus instances, allowing learners to familiarize themselves with appropriate usage.

This research also illustrates that material such as subtitles is essential in a parallel corpus to capture spoken elements in quasi-spoken texts. Therefore, such sources should be expanded not only by increasing the collection of subtitles but also by including film or drama scripts to provide more relevant data for examining interlingual correspondence in spoken language. Despite their weaknesses, subtitles have certain advantages that are often overlooked by some researchers (Charciarek, 2018, 2019), who cite only a few benefits (such as being a source of idioms for translation dictionaries) and ignore the usefulness of subtitles in analysing the features of spoken language.

Potential limitations of this study might arise from the analysis of data sourced from film subtitles. Although film subtitles strive to replicate spoken language, there is a lack of empirical evidence to assess their proximity to real spoken language. Consequently, their authenticity could be questioned. Hence, further research is necessary to gauge the extent of similarity between film subtitles and naturally occurring spontaneous informal conversations. This could be achieved through a comparative analysis, utilising a multi-dimensional space where any dataset can be compared against a model of the Czech language. A similar approach was proposed in previous studies that compared web-crawled and traditional corpora (Cvrček, Komrsková, et al., 2020) or elicited texts (Cvrček, Laubeová, et al., 2020a). By projecting film subtitles onto a multi-dimensional model of Czech, their prevalence within both spoken and non-spoken texts could be revealed.

The study substantially adds to our understanding of spoken expressions and discourse markers in Czech and their English, French, and Polish equivalents relevant to spoken registers. The results offer valuable contributions to teaching and learning speaking skills in Czech through corpus-based exercises. It is a first attempt to use corpus methods for familiarizing learners with spoken language using authentic Czech texts. This approach potentially shapes the further development of teaching methods, which have traditionally relied on prepared texts, and could considerably enhance the development of learners' spoken language. Furthermore, parallel data provide a wide array of translation equivalents; unlike printed dictionaries, they additionally offer contextual examples and rare, occasional translations that may be suitable only in specific situations. Traditional dictionaries typically provide the most frequent equivalents and often become outdated. The research provides a clear explanation of using corpus-based exercises in the classroom to foster learners' autonomy in developing fluency and achieving a more native-like performance in oral communication. It also contributes to previous studies conducted on the English language (Boulton, 2009; Huang, 2019; Şahin Kızıl & Savran, 2018) that do not use parallel corpora to identify translation equivalents in the learners' native languages, a crucial aspect for comprehending structures in a foreign language.

OPEN ACCESS

Acknowledgements

My thanks go to Petra Poukarová for fielding my questions about discourse markers and to Alexandr Rosen for his supportive comments.

This work was supported by the Cooperatio Program, research area Linguistics, and the project *Multilingual Lens: Investigating Large Text Corpora from Different Methodological Perspectives* (UNCE/24/SSH/009).

REFERENCES

Baker, P., & Ellece, S. (2011). *Key terms in discourse analysis*. London: Bloomsbury Publishing.

Bańczyk, Ł., Dybalska, R., & Vavřín, M. (2019). *InterCorp — Polish, Release 12 of 12 December 2019* [Corpus]. Prague: Institute of the Czech National Corpus, Charles University. www.korpus.cz

Barlow, M. (2000). Parallel texts in language teaching. In S. P. Botley, A. M. McEnery, & A. Wilson (Eds.), *Multilingual Corpora in Teaching an Research* (pp. 106–115). Amsterdam-Atlanta: Rodopi.

Bermel, N. (2014). Czech Diglossia: Dismantling or Dissolution? In J. Arokay, J. Gvozdanovic, & D. Miyajima (Eds.), *Divided Languages? Diglossia, Translation and the Rise of Modernity in Japan, China, and the Slavic World* (1st ed., pp. 21–37). Dordrecht: Springer International Publishing.

Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*, *21*(1), 37–54.

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning*, *67*(2), 348–393. https://doi.org/10.1111/lang.12224

Bulejčíková, P. (2015). *Problematika spisovnosti se zřetelem k výuce češtiny jako cizího jazyka [Standard Language with Regard to Teaching Czech as a Foreign Language]* [Dissertation, Charles University]. Charles University, Prague. https://dspace.cuni.cz/handle/20.500.11956/63386

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427.

Čermáková, A., Jílková, L., Komrsková, Z., Kopřivová, M., & Poukarová, P. (2019). Diskurzní markery. In J. Hoffmannová, J. Homoláč, & K. Mrázková (Eds.), *Syntax mluvené češtiny* (pp. 244–351). Praha: Academia.

Charciarek, A. (2018). Možnosti využití korpusu InterCorp v česko-polské překladové lexikografii. *Časopis pro moderní filologii*, *100*(2), 206–222.

Charciarek, A. (2019). Využití paralelního korpusu v translatologii (na základě česko-polského InterCorpu). *Bohemistyka*, *XIX*(2), 194–216. https://doi.org/10.14746/bo.2019.2.5

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A. J., & Benko, V. (2020). Comparing web-crawled and traditional corpora. *Language Resources and Evaluation*, *54*, 713–745. https://doi.org/10.1007/s10579-020-09487-4

Cvrček, V., Laubeová, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2020a). Author and register as sources of variation: A corpus-based study using elicited texts. *International Journal of Corpus Linguistics*, *25*(4), 461–488. https://doi.org/10.1075/ijcl.19020.cvr

Cvrček, V., Laubeová, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2020b). *Registry v češtině*. Praha: Nakladatelství Lidové noviny.

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal*, *58*(4), 363–374.

Holá, L. (2019). Běžně mluvená čeština ve výuce češtiny jako cizího jazyka. In M. Nekula & K. Šichová (Eds.), *Variety češtiny a čeština jako cizí jazyk* (pp. 107–127). Praha: Akropolis.

Hrdlička, M. (2019). Spisovná a obecná čeština ve výuce cizinců. In M. Nekula & K. Šichová (Eds.), *Variety češtiny a čeština jako cizí jazyk* (pp. 85–106). Praha: Akropolis.

Huang, L. (2019). A corpus-based exploration of the discourse marker well in spoken interlanguage. *Language and Speech*, *62*(3), 570–593. https://doi.org/10.1177/0023830918798863

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *Classroom Concordancing: ELR Journal*, (4), 1–16.

Klégr, A., Kubánek, M., Malá, M., Rohrauer, L., Šaldová, P., & Vavřín, M. (2019). *InterCorp — English, Release 12 of 12 December 2019* [Corpus]. Prague: Institute of the Czech National Corpus, Charles University. www.korpus.cz

Komrsková, Z., Kopřivová, M., Lukeš, D., Poukarová, P., & Goláňová, H. (2017). New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Journal of Linguistics/Jazykovedný Časopis*, *68*(2), 219–228. https://doi.org/10.1515/jazcas-2017-0031

Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., & Škarpová, M. (2017). *ORTOFON v1: Korpus neformální mluvené češtiny s víceúrovňovým přepisem*. [Corpus] Praha: Ústav Českého národního korpusu FF UK. www.korpus.cz

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Škrabal, M., Truneček, P., Vondřička, P., Zasina, A. J. (2015). *SYN2015: Reprezentativní korpus psané češtiny* [Corpus]. Praha: Ústav Českého národního korpusu FF UK. www.korpus.cz

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Škrabal, M., Truneček, P., Vondřička, P., Zasina, A. J. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2522–2528. Portorož: ELRA.

McCarthy, M., & Carter, R. (2001). Size Isn't Everything: Spoken English, Corpus, and the Classroom. *TESOL Quarterly*, *35*(2), 337–340. https://doi.org/10.2307/3587654

Nádvorníková, O., & Vavřín, M. (2019). *InterCorp — French, Release 12 of 12 December 2019* [Corpus]. Prague: Institute of the Czech National Corpus, Charles University. www.korpus.cz

Rosen, A., Vavřín, M., & Zasina, A. J. (2019a). *InterCorp — Czech, Release 12 of 12 December 2019* [Corpus]. Prague: Institute of the Czech National Corpus, Charles University. www.korpus.cz

Rosen, A., Vavřín, M., & Zasina, A. J. (2019b). *InterCorp, Release 12 of 12 December 2019* [Corpus]. Prague: Institute of the Czech National Corpus, Charles University. www.korpus.cz

Şahin Kızıl, A., & Savran, Z. (2018). The Integration of Corpus into EFL Speaking Instruction: A Study of Learner Perceptions. *International Online Journal of Education and Teaching*, *5*(2), 376–389.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511611841

Škrabal, M., Laubeová, Z., & Štěpánková, B. (2022). *Korpusové přístupy k české diglosii*. Praha: Nakladatelství Lidové noviny.

Škrabal, M., & Vavřín, M. (2017). Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii*, *99*(2), 245–260.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2007). Parallel corpora for medium density languages. In N. Nicoloy, R. Mitkov, G. Angelova, & K. Bontcheva (Eds.), *Recent Advances in Natural Language Processing IV* (pp. 247–258). Amsterdam/Philadelphia: John Benjamins Publishing.

Vavřín, M., & Rosen, A. (2015). *Treq* (2.0) [Computer software]. Praha: FF UK. https://treq.korpus.cz/

Vondřička, P. (2014). Aligning parallel texts with InterText. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1875–1879).

Reykjavik: European Language Resources Association (ELRA).

Walsh, S. (2010). What features of spoken and written corpora can be exploited in creating language teaching materials and syllabuses. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 333–344). London — New York: Routledge.

Zasina, A. J. (2023). *Korpusová cvičebnice pro studenty češtiny jako cizího jazyka*. Praha: Karolinum.

Adrian Jan Zasina
Institute of Czech and Deaf Studies
Faculty of Arts, Charles University
Na Příkopě 584/29
110 00 Praha 1
ORCID ID: 0000-0001-9348-5833
adrian.zasina@ff.cuni.cz