Charles University

Faculty of Arts

Department of English Language and ELT Methodology

Study programme: Philology (English Language)

Dissertation

Mgr. Denisa Šebestová

A contrastive description of English and Czech using the methodology of n-gram extraction

Kontrastivní popis angličtiny a češtiny s využitím metodologie n-gramů

dissertation supervisor: doc. PhDr. Markéta Malá, Ph.D.

dissertation consultant: PhDr. Jiří Milička, Ph.D.

2022

Thanks

My sincere thanks go to Markéta Malá for being my teacher, supervisor, reader and listener – and being exceptionally kind, patient and helpful in all those roles over many years.

I am much obliged to Jiří Milička for his immense help with the n-gram extraction software, and providing me with continuous technical as well as moral support.

For their advice and inspiring comments at various points, I am thankful to Gabriela Brůhová, Anna Čermáková, Viktor Elšík, Łukasz Grabowski, Nicholas Groom, Magdaléna Králová Zíková, Michal Láznička, Olga Nádvorníková, Lucie Steidlová, Ondřej Tichý, Adrian Jan Zasina, as well as to Michal Křen and Pavel Vondřička for technical assistance.

Thanks to the Anglo-Czech Educational Fund for enabling me to conduct research abroad.

Special thanks to Martin Sedláček for always having my back.

Last but not least, I am grateful to Jana Tolletová for turning me on to English; to my parents; and to all my colleagues from ÚAJD for being most kind and supportive.

I hereby declare that I have written this dissertation independently, using only the mentioned and duly cited sources and literature, and that the work has not been used in another university study programme or to obtain the same or another academic title.

Prohlašuji, že jsem disertační práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze 8. června 2022 Mgr. Denisa Šebestová, v.r.

Abstrakt

Tato disertační práce zkoumá frazeologické vzorce ve třech různých registrech (parlamentní debaty, noviny a dětská beletrie) v angličtině a češtině. Vyhledává a rozebírá rekurentní sekvence slov pomocí metody n-gramů. Cílem je popsat frazeologické vlastnosti každého registru a porovnat je mezijazykově. Práce také zkoumá možnosti přizpůsobení metody n-gramů se zřetelem k typologickým vlastnostem zkoumaných jazyků. Čeština v tomto ohledu klade na metodu vyšší nároky kvůli velké míře morfologické a slovosledné variability.

Práci tvoří tři případové studie, každá věnovaná jinému registru. První studie využívá a následně porovnává různé délky n-gramů. Předmětem zkoumání je malý korpus úzce specializovaného registru parlamentních debat. Studie dochází k závěru, že komplexního popisu registru lze dosáhnout nejlépe kombinací n-gramů různých délek. To ovšem představuje metodologický problém:, protože mezi n-gramy různých délek jsou četné překryvy, které znemožňují přesnou kvantifikaci a způsobují, že některé n-gramy jsou zastoupeny vícekrát. Studie popisuje různé funkce frazeologických vzorců a všímá si jejich role v organizaci diskursu. Druhá studie je věnována novinovým textům a zaměřuje se na n-gramy obsahující některou z předem zvolených předložek; předložky jsou ke zkoumání frazeologických vzorců vhodným východiskem díky tomu, že jsou frekventované a přispívají ke strukturování textu. N-gramy zde doplňuje rozbor kolokací, které odhalují evaluativní prosodie a sémantické preference frazeologických vzorců; výsledky poukazují k tomu, že novinové texty nejsou čistě informační. V poslední studii zkoumám frazeologické vyjádření času v beletrii pro děti. Východiskem jsou tentokrát n-gramy obsahující slovo, které spadá do předem vymezené sémantické oblasti časových významů. V této studii se pokouším s pomocí softwaru Engrammer (Milička 2019) překonat metodologické obtíže, které se objevily v předchozích studiích. Data jsou lemmatizována s ohledem na morfologickou variabilitu; v n-gramech je umožněna slovosledná variabilita a jsou slučovány n-gramy, které se částečně překrývají.

Ve všech třech studiích nesou frazeologické vzorce v češtině i angličtině podobné významy, které plynou z komunikačních potřeb daných konkrétním registrem. Mezijazykové rozdíly mezi vzorci jsou přisuzovány rozdílům typologickým (např. delší vzorce v angličtině jsou důsledkem její analytické povahy); jiné rozdíly jsou interpretovány ve světle možných kulturních rozdílů (např. povaha interakcí v parlamentních debatách). Práce dochází k závěru, že n-gramy jsou účinným prostředkem k odhalení frazeologických vzorců; ke zjištění, jak tyto vzorce fungují v kontextu je však optimální n-gramy kombinovat s dalšími metodami, jako např. přístup lexis-to-n-grams (n-gramy obsahující konkrétní slovo) nebo kolokace n-gramů. Optimální je metodu n-gramů přizpůsobit typologickým vlastnostem zkoumaných jazyků, a to zavedením lemmatizace, umožněním slovosledné variability uvnitř n-gramů či slučováním překrývajících se n-gramů.

klíčová slova: n-gramy, frazeologie, corpus-driven, registr, čeština, angličtina

Abstract

This dissertation examines phraseological patterns in three registers (parliamentary debates, newspaper reporting, children´s fiction) in English and Czech. It identifies and analyses recurrent word sequences through n-gram extraction, aiming to characterise the phraseology of each register and compare them cross-linguistically, while observing how the n-gram method can be adapted to accommodate for the typological properties of each language. Czech is particularly challenging in this respect due to its morphological and positional variability.

The dissertation comprises three case studies, each focussed on a different register. The first case study explores different n-gram lengths using a small corpus of a specialised register – parliamentary debates, suggesting that for a comprehensive register characterisation, different lengths should be combined. It notes the importance of discourse-structuring patterns and the problem of overlaps between n-grams. In the newspaper study, I extract n-grams containing prepositions – a convenient starting point given their frequency and involvement in text-structuring. N-grams are complemented with collocation analysis, revealing some evaluative prosodies and semantic preferences of patterns and suggesting that the newspaper register is not purely informational. Finally, I focus on the phraseological expression of temporal meanings in children´s fiction, extracting n-grams containing a word from a pre-defined semantic group. The methodological problems identified in the previous studies are addressed with the help of Engrammer software (Milička 2019): the data is lemmatised to control for morphological variability, flexible word order is enabled (shuffling) and overlapping n-grams are collapsed.

In all three studies, patterns are found to convey largely similar meanings in English and Czech, dictated by the communicative needs of the respective registers. The differences observed are attributed to language typology (e.g. longer English n-grams); others are discussed in the light of potential cultural differences (parliament interactions). N-grams are found to be efficient in identifying phraseological patterns; however, to examine how these patterns operate in context, n-grams are best combined with other methods (lexis-to-n-grams, collocation). Lemmatisation, shuffling and collapsing help refine the method and yield more informative results with respect to the typological properties of the languages compared.

keywords : n-grams, phraseology, corpus-driven, register, Czech, English

Contents

List of tables and figures

1 Introduction

1.1 Objectives & hypotheses

The present thesis is concerned with a phraseological analysis performed on English and Czech data representing different registers. It adopts a frequency-based perspective on phraseology, i.e. focussing on frequently recurring multi-word expressions. The chief motivation for the present study stems from the limited insight into Czech frequency-based phraseology based in a contrastive (cross-linguistic) framework: to date, there have been few studies comparing Czech and English from a distributional-phraseological perspective (Čermáková & Chlumská, 2017). Secondly, the study responds to previous research applying n-gram methodology to cross-linguistic data. As Granger (2013: 1) points out, despite the general awareness of the potential assets of the contrastive approach in phraseology, in practice it is rather underexploited as "systematic contrastive analyses of lexical bundles in different languages are very rare".

In corpus linguistics, two approaches are often distinguished: corpus-based as opposed to corpus-driven (Tognini-Bonelli, 2001, pp. 84–87). While in practice most corpus analyses combine both approaches to some degree, broadly speaking, corpus-based studies use corpus material to test and verify a preconceived hypothesis. Corpus-driven studies first identify recurring phenomena and trends in data, which in turn provide a starting point towards formulating a hypothesis (Altenberg, 1998; Tognini-Bonelli, 2001, pp. 84–87). This study is corpus-driven, adopting an exploratory perspective and using n-grams (i.e. continuous strings of n tokens found in corpus data) as the starting point. N-gram methodology seems well suited for this study since it allows for an automatic large-scale extraction of recurrent word sequences. In line with the frequency-based approach to phraseology (J. Ebeling & Ebeling, 2013; Gray & Biber, 2015), such recurring multi-word chunks are viewed here as "key components of language" (Colson, 2008). Moreover, recurrent word combinations have been found to perform a variety of specific functions in a text, helping to create and shape its structure; in Mahlberg's words (2013: 56) "frequent sequences of words have important discourse functions as textual building blocks". Importantly, n-grams have also been shown to be highly sensitive to register (e.g. Biber et al., 2004; Gries et al., 2011). Therefore, the n-gram method offers an opportunity to identify and examine the salient features of a specific register.

With all this in mind, three case studies have been conducted, each illustrating a different approach and use of the n-gram method, and each exploring a dataset representing a different register. The presumption is that since n-grams reflect register characteristics (Biber et al., 2004), they will likely pose different methodological problems in each study depending on the respective register. Hence, the n-gram method will be altered and adapted in each study based on the functional characteristics of the register at hand; furthermore, throughout the three studies, the aim is to gradually refine the n-gram method based on findings from the preceding studies. The registers for each study have been chosen and ordered with respect to the complexity of their communicative functions, which increases in each study: starting with parliamentary debates, a functionally highly specialised register; moving on to newspaper reporting, whose function is primarily informational (i.e. relatively broader than that of parliamentary debates, but still fairly homogeneous); and finally examining children´s fiction, characterised by a rich variety of functions (Hunt, 2005).

Along with these methodological adaptations, I aim to examine how the results are affected by modifying the basic variables of the n-gram search, namely n-gram length and internal variability (variable word order and/or open slots in n-grams). I also experiment with introducing lemmatisation to n-gram extraction. Each register is characterised through the lens of its most frequent recurrent word-combinations, focussing on their semantic and functional properties in order to examine what "propositional (semantic and grammatical)" (Altenberg, 1998, p. 121) and pragmatic functions they fulfil. This perspective was selected to compare phraseology between registers, but also to examine whether there are any cross-linguistic similarities in the phraseologies of individual registers. With regard to previous research, I expect this to be the case.

To summarise, my study aims to contribute to the contrastive description of Czech and English phraseology based on corpus data. It employs quantitative, corpus-driven (Tognini-Bonelli, 2001, pp. 84–87) n-gram methodology cross-linguistically in order to test its efficiency and applicability to English and Czech, attempting to overcome the issues generated by this language pair being typologically distant. N-grams are expected to highlight the typological characteristics of each language: Czech n-grams are likely to display a greater degree of variability in terms of word order and inflection; English ones to contain a large number of function words due to the analytical nature of English (cf. Gray & Biber, 2015, p. 136). As a result, the n-gram method is likely to be curbed when applied to Czech data; however, the cross-linguistic perspective will help as a diagnostic, making the limitations more explicit.

This study further aims to examine the phraseological characteristics of three selected registers and compare them between English and Czech, investigating each register in a separate case study. The n-gram method is adapted and refined for each case study with regard to the function of the given register.

The study thus has a methodological as well as analytical focus. My aim is to compare how the factors of typological characteristics and register determine the nature of frequent recurrent word combinations, and how they translate into phraseological differences between Czech and English. Overall, Czech is expected to lend itself less readily to an n-gram-based investigation, due to its flexible word-order and high degree of inflection.

1.2 Structure outline

The first chapter establishes the theoretical framework for the study. It introduces the key concepts in frequency-based phraseology, and discusses associated terminological and methodological issues. It outlines possible approaches to phraseology, focussing mainly on the study of phraseology through corpus methods, and explains the methodological rationale of the thesis. The chapter goes on to explain the motivation for adopting a contrastive approach, arguing for its potential assets in combination with a phraseological perspective. The following three chapters each present a case study. These studies all employ n-gram-based methodology to explore comparable Czech and English datasets; each uses a different cross-linguistic corpus pair representing a different register. The three registers were chosen so as to cover a wide scope of language functions: parliamentary debates, children's fiction, and newspaper texts. Detailed data descriptions and respective hypotheses are presented in the opening of each case study. While all case studies are n-gram-based, each employs a slightly different approach to the n-gram method, expanding on the methodological findings of the previous case studies.

Firstly, a pilot study is conducted on a small corpus of parliamentary debates. The design of this study is exploratory, aiming to compare what aspects of the register are revealed by n-grams of different lengths (between 2 and 10). The following case study deals with newspaper texts and is methodologically more complex. The focus is narrowed down to n-grams containing prepositions in order to identify prepositional patterns, which are considered a potentially illuminating point of departure for examining phraseology due to their frequency (Groom, 2010) and textual functions (Hunston, 2008); while also potentially interesting crosslinguistically due to the frequent lack of correspondence between dictionary translation equivalents (Klégr & Malá, 2009). The n-gram search here is further complemented by a collocation analysis, to examine how the prepositional patterns operate in context and contribute to construing evaluative meanings (semantic prosody).

The third case study investigates children´s fiction, a register fulfilling a variety of functions (Hunt, 2005; Thompson & Sealey, 2007). The chapter attempts to deal with the previously identified methodological issues by using lemmatised data, allowing for flexible word order and empty slots in n-grams. A semantic restriction is applied to the n-gram method, searching for n-grams containing temporal expressions. This is informed by the nature of the register at hand, since time in children´s literature is claimed to be expressed in specific ways (Thompson & Sealey, 2007).

The final chapter summarises the results of the three case studies and discusses phraseological tendencies identified in the data, with a twofold aim: a) to list phraseological features characteristic of individual registers, observing to what extent they are found across the two languages; b) to identify recurrent language-specific phraseological features distinguishing English from Czech and vice versa. Finally, it evaluates the methodology employed and its suitability for examining the respective languages.

2 Theoretical background & methodological overview

2.1 Frequency-based phraseology: theoretical framework, assets

Since the 1980s, an extensive body of research has been based around the presumption that “a good portion of the language we use every day is composed of prefabricated expressions, rather than being strictly compositional” (Gray & Biber, 2015, p. 125); and, by extension, that “meaning […] reside[s] in multi-word units rather than single words” (Ebeling & Ebeling, 2013, p. 65), such units forming “complex linguistic gestalts” (Hanks, 2013, pp. 283–289).

Adopting this perspective on language has important theoretical implications – in Erman and Warren’s (2000, p. 29) words, it “makes it impossible to consider idioms and other multi-word combinations as marginal phenomena”; in fact, such multi-word combinations form a large portion of our language production.[1] Consequently, recent developments in phraseological studies have been marked by a shift away from areas of research such as the study of fixed idiomatic phrases, paremiology etc., as the notions of idiomaticity and phraseology have been expanded to cover more broadly conceived phraseological items, labelled as multi-word units or recurrent combinations of words. Some scholars adopt the term *phrasicon* or refer to “the phraseology of language” (cf. e.g. Ebeling & Hasselgård, 2015, p. 207). In the following paragraphs I provide a brief summary of the key developments in the field of phraseology and the state of the art, sketching the background for the key theoretical assumptions on which my study is based.

[1] Erman & Warren (2000) analysed the extent to which their corpus was composed of “prefabs”, or manifestations of the idiom principle (Sinclair, 1991). Their results suggest that on average, 55% of a text (be it spoken or written) will consist of prefabricated items.

2.1.1 Recent developments in phraseology

As pointed out by Ebeling and Hasselgård (2015, p. 207), phraseology has become firmly established as a subfield of linguistics only recently; "Granger and Paquot (2008: 27) link this late scientific recognition to the field's unruly terminology and its vast and apparently unlimited scope" (Ebeling & Hasselgård, 2015, p. 207). Whilst becoming recognised as a discipline *per se*, phraseology has also undergone a paradigmatic shift. Therefore, the distinction can be made between two approaches: "phraseological"[2] (termed taxonomic[3] by Groom, 2017), as opposed to "frequency-based" (also probabilistic[4] (Groom, 2017); distributional (Granger & Paquot, 2008); or data-driven (Granger & Meunier, 2008)).

Initially, the bulk of research in the field of phraseology originated in Russian and German linguistics (e.g. an authoritative study on the classification of idioms by Vinogradov (1946; cited in Colson, 2008, p. 192) and proved highly influential: "the flow of ideas in phraseology since the late 1960s has been almost entirely from East to West" (Cowie, 1998a, p. 209). In the Anglophone context, research with a thorough focus on phraseology only started to appear systematically in the 1980s (Cowie, 1998a, p. 218). At this point, lexicographers tended to favour the taxonomic approach, focussing on developing categorisations for idioms. On the taxonomic approach, idioms are defined based on criteria such as fixedness and opacity / non-compositionality of meaning (Čermák, 2017, p. 132), and juxtaposed with free word combinations. Over time, these criteria have proven problematic and difficult to employ in practice for several reasons. Above all, it became clear that speakers' perceptions of meaning

[2] As the wording "phraseological phraseology" seems rather clumsy, I will be using Groom's term *taxonomic*.

[3] "Because main interest is in developing and working with formal taxonomies of phraseological units" (Groom, 2017).

[4] Groom introduces the term *probabilistic* with reference to Sinclair's observation that "There are virtually no impossible collocations, but some are much more likely than others" (Sinclair, 1966: 411, cited in Groom, 2017).

opacity vary greatly, making this criterion idiosyncratic and highly unreliable (Cowie, 1998a, p. 215). Moreover, idiomatic expressions encompass a vast variety of formally diverse units. As described by Altenberg (1998, p. 122): "more or less conventionalized building blocks [...] used as convenient routines [...] come in all forms and sizes, from complete utterances to short snatches of words, and they display varying degrees of flexibility."[5]

With the advent of corpus linguistics, the growing amount of available authentic data made the notion of a clear-cut classification of idiomatic expressions even more problematic, as numerous corpus analyses have indicated that there is a vast number of words which "customarily co-occur" (Kjellmer, 1991, p. 112), including discontinuous units – collocations, which display varying degrees of fixedness or flexibility. This tendency towards "customary co-occurrence" proves highly widespread in language, prompting linguists to extend the scope of their view to encompass a range of more or less idiomatic types of units, covering a cline from fixed and opaque idioms to collocations.[6] In his study of recurrent word combinations in spoken English, Altenberg (1998, p. 106) concludes that:

> [E]ven if they are *not fully lexicalized* expressions, they represent *conventional* ways of conveying specific pragmatic meanings. It should be added, however, that -- unlike most idioms -- none of the expressions serves as a unique expression of its particular pragmatic function. Rather, as the grouping suggests, each function seems to be served by a set of expressions. (my emphasis).

[5] Also, the importance of phraseology accounts for the frequency of some items: to use Summers' observation (1996: 262-263, cited in Stubbs, 2002, p. 227), "some of the most frequent words in the language [... are] not frequent by virtue of their single word uses [...] but because they often occur in so many set phrases or chunks".

[6] Reflecting this broadly conceived phraseological approach, Erman (2014) even goes so far as to entitle her article "There is no such thing as a free combination".

Idiomaticity here is redefined in that the criterion of *conventionality* (or *customariness*) is favoured over that of *fixedness*.

More broadly speaking, the establishment and growth of frequency-based phraseology was rooted in the development of usage-based approaches to language, characterised by Tomasello (2000, p. 61) as follows:

> In usage-based models of language [...] all things flow from the actual usage events in which people communicate linguistically with one another. The linguistic skills that a person possesses at any given moment [...] result from her accumulated experience with language across the totality of usage events in her life.

In other words, linguistic knowledge is believed to be an abstraction based on linguistic usage. Consequently, this view questions the notion of a clear distinction between phraseological and non-phraseological word combinations (see also Groom, 2017), as well as the traditional distinction between grammar and lexicon.

The corpus-linguistic perspective was further supported by work in cognitive linguistics, particularly construction grammar (CxG),[7] whose analyses of authentic data suggested that the phenomenon of idiomaticity[8] plays a crucial role in language, and is by no means limited to fixed idioms only. To cite just one influential study, Fillmore et al. (1988, p. 501) conclude that "the realm of idiomaticity in a language includes a great deal that is

[7] The CxG framework relates to the corpus-linguistic approach to phraseological units in that the typology of constructions covers a continuum between fixed, "schematic" and semantically non-compositional combinations (such as taxonomic idioms, set phrases), and open, compositional ones (Fried, 2013, pp. 2–4).

[8] Compared to the defining criteria of taxonomic idioms, the notion of *idiomaticity* is here revisited accordingly; Fillmore et al. (1988, p. 504) suggest that an idiomatic item "is assigned an interpretation by the speech community but [...] somebody who merely knew the grammar and the vocabulary of the language could not, by virtue of that knowledge alone, know (i) how to say it, or (ii) what it means, or (iii) whether it is a conventional thing to say".

productive, highly structured, and worthy of serious grammatical investigation".

In summary, the expansion of the scope of phraseology has been motivated by the growing awareness of the importance of multi-word items in language: "current research demonstrates that phraseology in the broad sense is one of the key components of language and is probably universal" (Colson, 2008, p. 192). In line with this, Cowie's (1998b, p. 3168) definition of phraseology is relatively broad, disregarding traditional taxonomic criteria (fixedness, opacity) altogether: "'the study of the structure, meaning and use of word combinations".

To illustrate how the notion of phraseology has developed, we may consider the definition of *idiom* by Aarts et al. (1994, pp. 204–205) (interestingly they do not include a separate entry on *phraseology*):

> idiom: 1. A string of (more or less) fixed words having a meaning that is not deducible from the meanings of the individual words, e.g. *over the moon; under the weather* [...] Some of these phrases allow no alteration, except when used facetiously (**over the stars*, **kick the pail*); others allow some changes (*up to my/his/her/their, etc. eyes in work*).

This is largely in line with the taxonomic phraseological criterion of non-compositionality. Aarts et al. (1994, pp. 204–205) do point out, however, that "in some cases there is no very clear distinction between idiom, collocation, and fixed phrase." Their other definition of *idiom* is more tentative and inclusive, marking the ongoing shift towards a distributional approach (ibid.: 205):

> 2. A phrase that is fairly fixed (not necessarily with opaque meaning) but which shows, or appears to show, some grammatical irregularity, e.g. *these sort of people, come to think of it*. It is not unusual to find phrases such as *by car, on foot, in prison* (i.e. consisting of a

preposition with a normally countable noun, but without an article) described as idioms.

Notably, the criterion of non-compositionality is missing altogether from the latter definition.

The onset of frequency-based approaches to phraseology was marked among others by Pawley and Syder's (1983) seminal study. They distinguish between two types of language items: "memorised sequences" are semantically compositional and speaker-specific, i.e. each speaker possesses their own repertory . "Lexicalised sentence stems", on the other hand, are characterised as institutionalised, shared by speakers, and semantically non-compositional. They can correspond to various structural units, up to whole sentences (Pawley & Syder, 1983, p. 209). Taxonomic idioms would thus fall under lexicalised stems; although notably the vast majority of lexicalised stems are "literal" (transparent) in meaning, i.e. would not classify as idioms in the taxonomic sense. Pawley and Syder claim there may well be hundreds of thousands of such stems (ibid.). The formulation of this idea represents a crucial turning point, as earlier phraseology was centred around idioms in the taxonomic sense, which were considered somewhat peripheral elements.

To sum up, taxonomic phraseology typically works top-down: from a preconceived definition of an idiom towards identifying examples of its usage, describing and classifying them. By contrast, frequency-based (or indeed usage-based) phraseology proceeds "bottom-up", focussing on *parole* in Saussure´s terms (de Saussure et al., 2011): examining usage in order to formulate generalisations based on the tendencies observed in authentic linguistic data. In terms of its theoretical assumptions, "phraseology in the broad sense [*i.e. frequency-based*] meets the criteria of 'polylexicality' and 'fixedness', whereas phraseology in the narrow sense [*taxonomic*] requires the additional criterion of 'idiomaticity' (Colson, 2008, p. 193; my terms added in italics).

Notably, frequency-based phraseology is a dynamic field whose theoretical background is far from cast in stone – to date there is no unified theoretical framework (Colson, 2008, p. 194). Also, the exact scope of phraseology seems difficult to pinpoint (Altenberg, 1998, p. 100; Groom, 2017); approaches are often combined to explore the phraseology of a language from a variety of perspectives. There are related frameworks and approaches whose theoretical outsets are relevant and applicable to frequency-based phraseology, and which often work in parallel with it, such as CxG, cognitive linguistics (especially research on metaphor and figurative language, cf. Lindquist & Levin, 2008), and also e.g. pattern grammar (Hunston & Francis, 2000).

2.1.2 Frequency-based phraseology: key ideas

Sinclair (1991, pp. 109–111) introduced the notion of two competing principles shaping language production: the open-choice principle allows language users to combine language items creatively, as long as the resulting combinations are grammatical and meaningful. Thus, at any point, the user is free to choose from a vast number of possible means of expression. Clearly, the open-choice principle alone does not pose sufficient restrictions to ensure that fully acceptable and comprehensible results are produced; nor can it account for the existence of "preferred ways of saying things" (Altenberg, 1998, p. 122). Some choices are restricted by the logic of extralinguistic reality; yet others are determined by register. A further – and crucial – limitation to the open-choice principle, argues Sinclair, is provided by the principle of idiom: the user draws on "a large number of semi-preconstructed phrases that constitute single choices", and which function as a single semantic unit (1991, p. 110). Altenberg (1998, p. 115) arrives at a similar conclusion:

> [R]ecurrent […] sequences […] can be regarded as a series of *overlapping and interlocking options* that are utilized again and again by speakers in ongoing discourse. These sequences do not

> necessarily constitute phraseological units, but [we may] regard them as *more or less prefabricated or routinized building blocks that are at the speaker's disposal* in the production of discourse [...] (my emphasis).

This is in line with Sinclair's idea of the open-choice vs. idiom principles, with the added note that there is a "fuzzy boundary between fully lexicalized units and free expressions" (Altenberg, 1998, p. 118).

To summarise, during production, speakers exploit both their knowledge of recurrent, more-or-less fixed sequences of words from their mental lexicon, as well as an awareness of their combinatorics.[9]

Arguably, Sinclair's notion of idiom and open-choice principles is compatible with Hoey's (2005) theory of lexical priming. Hoey points out that most theories preserve the lexis vs. grammar distinction, their crucial criterion being whether a sentence is grammatical. He makes a case for the additional criterion of naturalness, which results from the notion of lexical priming. During acquisition, every word (or word sequence) becomes "cumulatively loaded with the contexts and co-texts" (Hoey, 2005, p. 8) in which it occurs. As a result, the word becomes primed to co-occur with such frequently recurring contexts: this priming shapes its collocations, but also colligations (co-occurrence with particular grammatical functions), semantic and pragmatic associations. Thus, priming has implications for grammar and textual organisation (ibid.).[10] Consequently, the knowledge of a word's meaning comprises the

[9] Remarkably, as pointed out by Groom, 2017, although Sinclair's conception of the idiom and open-choice principles considerably expands the scope of phraseology, it still retains the distinction of grammar as opposed to lexicon. However, Altenberg (1998, p. 100) recognises (based on Pawley & Syder, 1983) that "the existence of a large number of more or less prefabricated expressions in language blurs the distinction between lexicon and grammar". Likewise, construction grammar (e.g. Fried, 2013) abandons this clear-cut distinction, arguing that it does not reflect the representation of language in the human mind (Fried, 2013, p. 2).

[10] Interestingly, priming does not work in a straightforward systematic fashion in that it is not necessarily shared among synonyms or cohyponyms (Hoey, 2005, p. 13).

knowledge of its contexts. This may be linked to phraseological competence, which is found to be a crucial component of overall linguistic competence (Howarth, 1998; Paquot, 2018). The degree of phraseological competence is also an important criterion of determining L2 fluency, distinguishing native speakers from non-native learners (e.g. Granger & Bestgen, 2014; Hasselgård, 2019).

Considering its theoretical implications, lexical priming can be viewed as “the driving force behind language use, structure, and even language change”, as primings can evolve over time (Hoey, 2005, p. 12). Hoey thus develops a priming-based theory of language, concluding that "lexis is complexly and systematically structured and [...] grammar is the outcome of this lexical structure" (2005, p. 1). He builds upon the idea (based on Giddens 1979; cited in Hoey, 2005: 8) that “each individual action reproduces the structure and the structure shapes the individual action” (Hoey, 2005, p. 8). Hoey’s idea of grammar is readily compatible with the corpus-linguistic view; as Mahlberg (2013, p. 4) remarks, “in a corpus theoretical approach, a grammar is seen as a set of generalisations about the behaviour of words in texts”.

2.2 Terminology: n-grams, bundles, sequences

Phraseology (in the sense of *phrasicon*) encompasses a range of means of expression, including “collocations, phrasal verbs, compounds, idioms, speech formulae, and lexical bundles” (Paquot, 2018, p. 31). As mentioned by Stubbs, researchers exploring the phraseology of a language often have to face methodological conundrums “since there are severe problems in defining phrasal units in corpora, it is difficult to know what to count” (Stubbs, 2002, p. 215). In my study I will be working with n-grams and units derived from them; some other options are briefly discussed below.

The terminology employed in frequency-based phraseology research varies widely. Different scholars refer to “multi-word prefabricated expressions" (Biber & Reppen, 2015) by different terms. Essentially, we may differentiate between two types of units (after Stubbs, 2002, p. 230):

a) simple contiguous strings of words retrieved from corpus data, which may or may not correspond to complete linguistic units: cf. the fragments *and what he; on the* as opposed to the complete *on the other hand* (examples taken from case study I, this volume).In this thesis such strings will be termed *n-grams* after Lindquist and Levin (2008). Other terms for these strings include e.g. "clusters" (Mahlberg, 2013; Scott (1997:41; cited in Stubbs, 2002, p. 230), "lexical bundles" (Biber et al., 2004), "recurrent combinations" (Altenberg, 1998), "multiword combinations/items/sequences" (Webb, 2019), "multi-word units" (Granger & Meunier, 2008), "formulaic sequences" (Hyland, 2008), "phraseological sequences" (Vašků et al., 2019); or "stems" (Granger, 2014; Pawley & Syder, 1983).

b) structured and semantically complete word combinations, which may be interrupted (variable to a certain extent) or not. In the present volume they are termed "patterns" (Lindquist & Levin, 2008).[11] Other equivalent terms from the literature include "prefabricated patterns" or "prefabs" (Erman & Warren, 2000), "set phrases" (e.g. Colson, 2008).

Additionally, there are more general terms, including "comings-together-of-words" (Palmer 1933, cited in Cowie, 1998a, p. 211) or "wordcombinations" (Cowie, 1998a).[12] As pointed out by Groom (2017), this array of terms in usage does not reflect any vast theoretical disagreement within the field; rather, the choice of a particular term is motivated by the methodological approach adopted. Let us illustrate this using several examples pertaining to continuous strings:

[11] Specifically, Hunston and Francis (2000, p. 3) conceptualise "pattern" as the "attendant phraseology" of a word.

[12] A related notion is collocation – more on that in the following chapter.

Altenberg (1998, p. 101) refers to "recurrent word-combinations", specified as "continuous strings of words occurring more than once in identical form". The requirement of identical form makes this definition rather narrow; it proves inconvenienent to this study due to the inflectional character of Czech.

Biber et al. (1999: chapter 13) introduce "lexical bundles", defined as the most frequently recurring word combinations in a register, occurring at least ten times per million words (ibid.: 989). Biber et al. (2004) then categorise lexical bundles into three groups based on their function: referential bundles, referring to the situational or textual context; discourse organizers, expressing relationships within the discourse; stance bundles, expressing the speaker's stance or degree of certainty (conveying epistemic modal meaning). Another frequently used term is "cluster", which is not required to meet any given statistical or frequency criteria (Čermáková & Chlumská, 2016, p. 167).

In this thesis I adhere to the terms n-gram and pattern after Lindquist and Levin (2008, p. 144), distinguish between *n-grams* – "recurring strings (with or without linguistic integrity)" as opposed to *patterns* – "meaningful, linguistically structured recurring sequences of words". I view n-grams as an initial step aiming to automatically identify recurrent sequences in a corpus. Within these n-grams, I will be subsequently looking for *patterns* – items characterised by carrying a distinct meaning and conveying a particular function in the context. N-grams may correspond to patterns, or form fragments of patterns – hence they need to be examined in context.

2.3 Frequency-based phraseology: methods & issues

2.3.1 Identification of multi-word units

As pointed out by Čermáková and Chlumská (2016, p. 169), multi-word units can be identified using one of two broadly defined methods: either through the automatic extraction of n-grams (adopting a maximally corpus-driven approach), or using pre-selected lexical units as a basis and identifying patterning around them. In the same vein, Stubbs (2002) compares and explores the potential uses of collocations ("habitual co-occurrences") and n-grams ("chains") to examine the frequency of word combinations.

Collocations are a syntagmatic unit. A syntagma involves a dependency relation; accordingly, collocations are typically described in terms of a node and its collocate (Stubbs, 2002, p. 216). However, as pointed out by Sinclair et al. (2004, p. 10), this should not imply a hierarchical relationship between the two units - rather, one can choose to view either part of a collocation as either the node or the collocate.

Stubbs points out that much work on collocation falls under two general trends: either in-depth studies focussing on a limited dataset, often "over-emphasizing idiosyncratic cases" (2002, p. 217); or very broadly conceived large-scale studies whose point of departure is the importance and omnipresence of collocation, but whose results actually provide fairly little quantitative insight into it. He suggests combining these two viewpoints to achieve the best of both worlds and stresses the importance of using an extensive and balanced dataset: to control for the potential overrepresentation of idiosyncratic items, as well as provide a robust basis for quantitative conclusions and generalisations (2002, pp. 217–218).

According to Nesselhauf (2004, p. 17), an important variable in the different views of collocations is their defining criterion: essentially this is either simple frequency of recurrence, or statistical significance of the words´ co-occurrence, i.e. collocation strength, identified through statistical (association) measures. An influential definition of collocation is one based around probability: a combination of words which "co-occur more often than their respective frequencies and the length of text in which they appear would predict" (Jones & Sinclair 1974: 19; cited in Nesselhauf, 2004, p. 8). In a related vein, Stubbs views collocation strength as the degree of “attraction” between node and collocate: it can be measured by counting the ratio between a) the number of node-collocate co-occurrences and b) all node occurrences (2002, p. 221).

N-grams on the other hand are conceived more loosely than collocations, further from the usual notion of a syntagmatic combination; this is due to the fact that upon the extraction of an n-gram, the retrieved sequence may or may not correspond to a phrasal structure (Stubbs, 2002, p. 230). One drawback of the n-gram method is that the automatic large-scale extraction may result in a bulk of expressions "of little phraseological interest" (Altenberg, 1998, p. 102), such as fragmentary function word sequences (e.g. in English, frequent bigrams tend to consist of a preposition and article, cf. case study I). Still n-gram lists may be a "useful starting point" (ibid.), although the researcher needs to approach the retrieved data selectively, imposing some limits on them. The criteria for such limits are admittedly “to a large extent arbitrary" (ibid.); an important factor may be simply the feasibility of the task at hand. In practice, the “selection” is usually carried out through limiting the length of n-grams extracted, and determining a cut-off point.

2.3.2 N-gram length

As a rule, the optimum n-gram length for the purposes of a given analysis needs to be decided *ad hoc* for a given study. There are several interacting factors which come into play:

a) The research question of the particular analysis: the optimum length is determined by the sort of patterning one aims to reveal. For instance, in a study focussing on prepositional patterns, bigrams may be a viable and efficient option. Forensic authorship attribution studies provide some interesting insights into the issue of n-gram length. Coulthard (2004) has found that the longer a word sequence, the less likely it will be repeated by two different writers; the most reliable results in this authorship attribution study were yielded by 6-8-grams. This is not surprising: as a rule, a greater n-gram length implies more potential variability, in turn giving rise to a greater range of possible combinations (Cvrček & Václavík, 2015, p. 34). However, obviously the search for long sequences alone is only part of the story; the longer the sequence, the lower its overall frequency – even within texts by the same author (Wright, 2017, p. 220). Wright takes up an exploratory approach, extracting n-grams between 2 and 6 words long, finding that "the longer n-grams of five and six words in length outperform the shorter ones when attributing smaller samples, but the reverse is the case for larger samples" (Wright, 2017, p. 222). Overall, best results (i.e. the most accurate attributions) were achieved with 4-grams, followed by 5-grams. Also, the success rate of each n-gram length differed depending on the author, suggesting that different language users repeat different lengths the most frequently (Wright, 2017, p. 233).

b) Corpus design should inform the selection of n-gram length (Gray & Biber, 2015, p. 137). In particular, it is the register represented by a given dataset which has a major impact on the nature as well as length of multi-word items extracted from it (Biber et al., 2004; Gries et al., 2011): although we may determine a tentative range of the most frequent n-gram lengths, still different registers may show a preference for recurrent sequences of different lengths. Some types of data even manifest a relatively high frequency of long grams (see case study I). Further, some registers prove more repetitive than others. This may result in a lower number of n-gram types overall but possibly higher relative frequencies of n-gram tokens.[13] Another important factor is corpus size: the smaller the corpus, the fewer long grams it will contain (Čermáková & Chlumská, 2016, p. 168).

c) The typological properties of the language in question: e.g. English is predominantly analytical, employing a wide range of high-frequency

[13] By n-gram type I mean a particular recurrent sequence identified by an n-gram search; its individual occurrences are n-gram tokens. Cf. case study II.

function words (determiners, auxiliaries, prepositions), which translate into the most frequent n-grams.

Recommendations regarding the optimum length of n-grams vary between different researchers and studies. For English, Biber et al. (1999: Chapter 13) recommend 3-grams, noting that longer sequences (such as 4-6-grams) tend to be more idiomatic. As regards Czech data, according to Cvrček and Václavík's results (2015), on average 4-grams cover the greatest variety of word combinations.

The issue of n-gram length becomes all the more complex in cross-linguistic studies. As pointed out by Hasselgård (2019), when n-grams are employed to compare an analytical language with an inflectional one (such as English with Czech), it should be borne in mind that multi-word units in the inflectional language will usually be shorter than their counterparts in the analytical one. Also, the cross-linguistic correspondences may be marked by structural differences, further complicating the comparison (Granger 2014). Čermáková and Chlumská (2016) opt for 3-, 4- and 5-grams while comparing Czech and English; they argue that to date, this range of lengths has turned out the most productive (noting, however, that most studies applied it to English data only). Thus, the applicability of different lengths needs to be tested separately for each language (and language pair).

To determine the optimum legth in a cross-linguistic analysis, the *Corpus Calculator*'s n-gram correspondence tool uses a model based around "the equivalent number of types in the languages compared" (Cvrček, 2021). The model operates on two parameters: n-gram length and frequency minimum, "since these are the main factors influencing the number of types" (ibid.). The user supplies the model with the chosen n-gram length and frequency threshold used in one language, and the model returns recommended average values of both these parameters in the other language. Based on this recommendation, the researcher should then optimize the n-gram length, or use a combination of 2 n-gram lengths, extracted in a particular ratio. For instance, if the optimum length in language 2 is 2.34-grams, the user is instructed to mix 66 % of 2-grams with 34 % of 3-grams (Cvrček, 2021). If we plug in the following parameters: [L1 = Czech, L1 n-gram size = 4, L1 frequency threshold = 10, L2 = English], the tool computes the corresponding English parameters thus: [L2 n-gram size = 5.4, L2 freq threshold = 11.4]. The results are illustrated by the plot in fig. 1. In practice, our English sample would need to consist of 60% of 5-grams and 40% of 6-grams, meeting the minimal frequency of 11, to achieve comparability with Czech 4-grams whose minimum frequency was 10 (Cvrček, 2021). The dashed line shows the resulting correspondence between the two datasets once the parameters for English are adjusted.

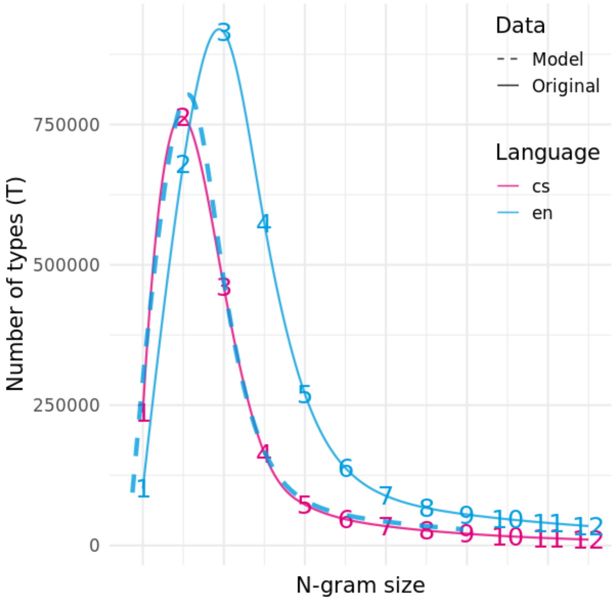


Figure 1. Recommended corresponding n-gram sizes for Czech (4-grams, freq > 10) and English (5.4 grams, freq > 11.4) as per *Calc* (Cvrček, 2021)

However, it should be added that the suitable n-gram length may be best determined experimentally through comparing the results yielded by different n-gram lengths (cf. case study I).

2.3.3 Open slot or not: phrase frames, *skipgrams*, *conc-grams*

As Altenberg (1998, p. 121) points out, word combinations are normally not entirely fixed, whether semantically or grammatically. Multi-word units vary in the extent to which they are formally fixed, ranging from completely frozen taxonomic idioms to highly flexible "phrase frames" – customary combinations featuring a variable unit in a given position (Fletcher, 2002). Structurally speaking, variability may manifest itself on various language levels: morphological, lexical, as well as word-order variability. Morphological and lexical variability will be discussed in more detail later as it constitutes a major issue when comparing Czech and English. Regarding lexical variability in frame-like constructions, one possible solution is to search for phrase frames, i.e. n-grams in which one position is variable, while the rest remains fixed (Fletcher, 2002).

Another approach is suggested by Cheng et al. (2006). They start from the term skipgram (essentially synonymous with Fletcher´s (2002) phrase frame), referring to "non-contiguous word associations" while also encompassing contiguous n-grams (introduced by Wilks 2005, cited in Cheng et al., 2006, p. 412). Searching for skipgrams allows us to identify contiguous as well as non-contiguous combinations; the latter would go unnoticed by a simple n-gram search (Cheng et al., 2006, p. 412). However, a major drawback to skipgrams is that they do not capture positional (word-order) variation. Cheng et al. therefore introduce the concept of *concgram*, covering both non-contiguous sequences and positional variation, while not being limited in terms of sequence size (2006, p. 413).[14] In my case study III, I will be using a similar method to *concgrams,* employing the custom-made software *Engrammer* (Milička, 2019).

[14] Cheng et al. (2006) have developed the ConcGram© software, specifically designed for this purpose. Importantly, ConcGram searches are fully data-driven, requiring no input from the researcher.

2.3.4 Frequency, statistical measures, cut-off points

During and after the retrieval of multi-word items from a dataset, whether through n-grams or other extraction methods, the researcher needs to establish criteria for the selection of items to be analysed, i.e. set a cut-off point: often, this is based on setting a required frequency threshold. The relationship between word frequency and corpus size is an important factor here. To quote Stubbs (2002, p. 227):

> Words are very unequal in frequency: a few words are very frequent, whereas most words are very rare. In a typical individual text or in small corpora of one million words or so, up to half the words will occur only once each.

Stubbs' observation can be linked back to Zipf's law 1, stating that a word's frequency multiplied by its rank remains constant. This implies an overall balance between lexical variety and word frequency: "consequently, any text will contain a very low number of high-frequency words and a majority of low-frequency words" (Cvrček, 2017; my translation).[15]

Collocation strength is another measure by which multi-words units can be sorted and selected for further analysis. According to Stubbs (2002, pp. 23–24), the "strength of attraction" between most collocations is between 2-10 %. Outliers on either extreme of this span are fairly rare; above 10 %, the stronger the attraction, the fewer collocations are found (ibid.). Although collocations can be identified based on simple frequency of co-occurrence, there are various association measures available, based on measuring the degree of association between node words, i.e. collocation strength. Association measures work with the frequency of co-

[15] "Důsledkem tohoto vztahu je fakt, že každý text obsahuje velmi malý počet slov frekventovaných a většinu slov málo frekventovaných" (Cvrček, 2017).

occurrence, compared to "expected frequency", i.e. how often the nodes would co-occur at random (Brezina, 2018, pp. 69–70). Often they also take into account the factor of corpus size. Widely used association measures include MI (mutual information), T-score, log-likelihood, Dice, logDice, or risk ratio. Importantly, each association measure favours collocations of a different kind, and thus is suited for different types of research questions (Brezina, 2018, pp. 66–67). Comparing the ranking of collocations supplied by several measures may therefore be revealing. Some measures (e.g. T-score) emphasise frequency of co-occurrence. Others, such as Dice, favour "exclusive" collocations, i.e. a word node's limited collocability is taken to be a major indicator of collocation strength. Brezina (2018, p. 74) provides a comprehensive overview of these tendencies or biases of individual measures, quoted here in fig. 2.

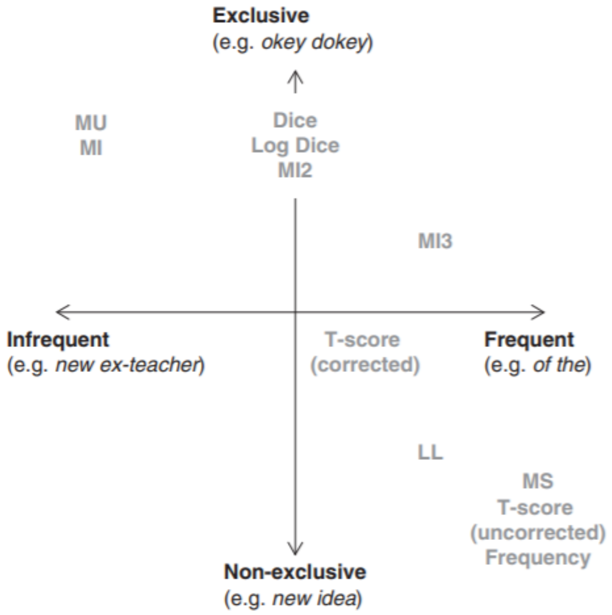


Figure 2. "Frequency and exclusivity scale" (Brezina, 2018, p. 74)

2.4 Contrastive/cross-linguistic approaches to phraseology

2.4.1 Contrastive[16] analyses: motivations, potential assets

As testified by the bulk of research dedicated to it, phraseology seems to operate in complex interplay with other language components such as semantics and syntax, while also being influenced by cultural factors (Colson 2008: 191). Therefore, combining a variety of approaches is optimal in order to achieve a comprehensive view of phraseology. In this regard, the cross-linguistic perspective is promising as it allows for an exploration of various constellations of linguistic as well as cultural factors interacting with phraseology. “A comparison between set phrases in two or more languages [is] of crucial importance for discovering the theoretical principles underlying phraseology, as well as its contextual use” (Colson 2008: 192). Furthermore, it provides valuable insight as “languages differ not just in the means of expression, but also in the *extent* to which particular meanings are conventionally expressed in natural discourse” (Johansson, 2007, p. 307, my emphasis). Finally, when combined with corpus methodology, a cross-linguistic approach provides the additional asset of enabling “a statistical analysis of the various categories of set phrases as well as a very reliable methodology” (Colson 2008: 191).

Possible exploitations of contrastive frequency-based phraseology include its application to typology studies. Phraseology may offer insight into language "universals", in the sense of identifying patterns underlying

[16]The meaning of the term *contrastive* may vary between researchers (Dobrovol’skij & Piirainen 2005; Korhonen 2007; both cited in Colson 2008: 193):
a) *contrastive* and *cross-linguistic* may be regarded as synonyms; or
b) *contrastive* is used in its narrow sense - systematic comparison of differences and similarities between the languages compared; or
c) *contrastive* in an even narrower sense, meaning only examining differences between languages.
In this study I use *contrastive* as in b) above, i.e. to refer to comparing languages with regard to both similarities and differences, and above all with a focus on function. I view *cross-linguistic* as a slightly broader term, as in “involving/employing several languages”.

the organisation of lexicon and finding out to what extent they correspond (or not) between languages. By extension, a cross-linguistic view may reveal how phraseology reflects the cultural characteristics of a particular language community. Another option is applying the findings in translatology, examining the language of translation and using the findings to inform translation practice (Colson 2008: 192).

2.4.2 Contrastive n-gram methodology: state-of-the-art

To date, n-gram methodology has proven to be a useful starting point for contrastive studies working with typologically related languages. When contrasting more distant language pairs such as English and Spanish, Norwegian, French, or Czech (Cortes 2008; Ebeling & Ebeling 2013; Hasselgård 2017a, b; Granger 2014; Čermáková & Chlumská 2016, 2017; Šebestová & Malá 2019, to name but a few), the methodology poses some problems rooted in typological non-correspondences between the languages compared. Observing such typological differences may prompt researchers to reconsider their theoretical framework and their notion of a multi-word unit.

Methodologically, the type of corpora used (either comparable or parallel)[17] predetermines the nature of a cross-linguistic study (Čermáková & Chlumská, 2016, p. 169). The two types of corpora may be combined, as in the ENPC (English Norwegian Parallel Corpus), compiled by Johansson (2007). The ENPC is designed to combine the advantages of both: the parallel corpus or originals and translations is a basis for identifying cross-linguistic correspondences, whilst the comparable corpus controls for potential cross-linguistic interference induced by the

[17] I will be using *parallel corpus* to mean a corpus of originals aligned with their translations.
On the other hand, I understand *comparable corpora* as collections of a similar size containing only original texts which are comparable in terms of register, distribution etc. (Cf. https://wiki.korpus.cz/doku.php/pojmy:paralelni)

translation process (Čermáková & Chlumská, 2016, p. 169).

I will now summarise the key findings and contributions of two selected recent studies comparing English with another language through n-grams which bear relevance to my own study.

Hasselgård (2017) compared English and Norwegian 2-4-grams extracted from the ENPC, restricted to original texts in both languages and focussing specifically on n-grams expressing temporal adverbial meanings. This study provides some important insights: it is valuable mainly in that it shows how n-gram methodology is able to highlight typological differences which would be difficult – if not downright impossible – to identify otherwise. The Norwegian data is shown to contain fewer recurrent n-grams overall, indicating that English may have a relatively stronger tendency towards recurrence. That being said, in Norwegian, temporal n-grams formed a larger part of all the n-grams identified compared to English. Also, Norwegian n-grams are shown to correspond to (fragments of) clauses more often than English ones (ibid.: 86). A possible implication of these results is that while some languages display more recurrence than others (i.e. typological properties are an important factor shaping the phraseology of a language), a language may employ phraseological means of expression to varying degrees in different semantic or functional areas (i.e. pointing towards a register-dependent distribution of phraseological means). These findings are therefore highly relevant for the purposes of this study.

Of equal importance are the limitations and difficulties which Hasselgård (2017) faced, as well as the limited success of this n-gram based analysis: for example, the hypothesis that English would be more nominal was not confirmed. This may suggest that while the n-gram method is suitable for an exploratory analysis and may reveal different tendencies in the languages compared, it may not be a very reliable means towards testing hypotheses about particular typological properties of the languages. However, Hasselgård states that using larger corpora and extending the size of the n-grams above 4 may be revealing.

Granger (2014) examined English and French data from comparable corpora, designed so as to allow for comparison across languages as well as registers – the data comprised parliamentary debates and newspaper editorials. She extracted a relatively wide range of n-gram lengths, between 3 and 7. The data was not lemmatised as lemmatisation was found problematic (a more detailed discussion of this follows in the next section). One interesting contribution of this study is the method of determining cut-off points. Granger selects them for each n-gram size separately, following the same procedure. For each n-gram length, she first extracts a set of frequency lists of n-gram types, changing the frequency threshold with each extraction. She proceeds to identify the cut-off point which yields approximately 0.10% of the total number of n-gram types. The aim is to ensure proportionality across different n-gram lengths (Granger, 2014, p. 63). Granger focuses on stems, defined by Altenberg (1998) as combinations of subject and verb, with optional pre-verbal thematic elements. In this study, French was expected to employ more lexical bundles overall than English. This tendency was clearly apparent in editorials, but rather inconclusive in parliamentary debates. Stems were found to be significantly more frequent in French than in English in editorials. Yet, the contrary was found in parliamentary debates (Granger, 2014, p. 64). Granger's results show that phraseological tendencies may differ markedly across languages as well as across registers. Therefore, she warns against overgeneralisations based on cross-linguistic bundle analyses (ibid.).

In summary, previous cross-linguistic research into n-grams/lexical bundles clearly indicates that typological properties and the register factor enter into a highly complex interplay, and it seems advisable to interpret results with a grain of salt, bearing in mind that depending on corpus design, their validity is likely limited to the particular registers explored.

2.4.3 Issues rooted in typological differences

The typological non-correspondences between Czech (inflectional with rich morphology, a wide repertory of endings; highly flexible word order) and English (analytical; with word order fixed due to its grammatical function) cause major issues. For instance, in Čermáková and Chlumská (2016) the English datasets of comparable size yielded hundreds of n-grams, whilst in the Czech data only tens of n-grams were identified altogether. This suggests that in order to ensure comparability of n-gram data across this language pair, precautions need to be taken; and the results for each language need to be examined separately as the cross-linguistic comparability may be severely limited.

2.4.4 Morphology

Being a predominantly inflectional language, Czech is characterised by wide morphological variety. This raises the question of lemmatising n-grams. Lemmatisation is helpful in that it helps to collapse n-grams which differ in terms of a particular morphological category, e.g. nouns differing in case, or verbs in person. This may be crucial because, as noted by Granger (2014), if one of the languages compared is considerably more inflectional than the other, this morphological richness may skew the results, producing a higher number of n-gram types in the more inflectional language, all of which in fact correspond to a single n-gram type in the more analytical language.[18] Finally, lemmatisation may be used to collapse n-grams containing the orthographic variants of a word.[19]

However, as both Granger (2014, p. 60) and Čermáková and Chlumská (2016, p. 170) point out, it is not advisable to apply lemmatisation across the

[18] For instance, Czech expresses additional morphological meanings such as gender, not grammaticalised in English.

[19] This may prove especially useful in diglossic situations, as in case of the two parallel language norms in Norwegian (Hasselgård, 2017, p. 79).

board. Some morphological categories reflect a semantic difference which we may want to preserve, such as e.g. the category of verbal mood. Thus, lemmatisation is best applied only selectively; or if n-grams are extracted from lemmatised data, the results should be interpreted with caution.

Another problem pertains to overgeneralisations potentially stemming from lemmatised n-grams. E.g. if we level out the category of noun case, the resulting lemmatised bigram *ON BÝT* [*HE BE*] may easily encompass individual sequences with entirely different meanings: *on byl [he was], on je [he is], jemu jsou [to-him are],* etc. (Cvrček & Václavík, 2015, p. 31). Thus, profitable as lemmatisation may be, it is admittedly no silver bullet for the cross-linguistic disparities resulting from different degrees of morphological complexity.

2.4.5 Word order

English is well-suited for the n-gram methodology as its word-order is relatively fixed, due to its grammatical function: a word's position in the sentence or phrase indicates its syntactic function (Dušková, 2015, p. 14). By contrast, in Czech, the use of n-grams is considerably limited by the high degree of word-order variability (Uhlířová & Kučerová, 2017). A possible way of controlling for this typological feature of of Czech is searching for positionally variable n-grams ("shuffle-grams"): e.g. the 4-gram hits *jednoho dne se ('one day se-reflexive'), se jednoho dne ('se-reflexive one day')* (cf. chapter 3 on children´s literature) would all be subsumed under one "shuffle-gram".

Furthermore, for word-order variability to be captured by n-grams, empty slots may be allowed within the n-gram (cf. "skipgrams" mentioned above, Cheng et al., 2006). Merging n-grams which are identical except for an optional element, such as an intensifier, filler, discourse marker etc., seems desirable (cf. *jednoho dne se, jednoho <u>krásného</u> dne se*).

2.4.6 Register characteristics as a factor

As noted by Biber and Conrad (2009, p. 6) "linguistic features [...] tend to occur in a register because they are particularly well suited to the purposes and situational context of the register".[20] Within a particular register, speakers choose the language they use with regard to their communicative purpose (Biber et al., 1999, p. 21).

As testified by a number of studies cited above, n-grams/multi-word expressions prove highly register-specific (Biber et al., 2004; Granger, 2014). Moreover, Johansson suggested the importance of "extend[ing] contrastive studies by taking into account the variation across registers within languages" (Johansson, 2007, p. 304). Therefore, register should be a principal defining criterion when compiling or selecting corpora for an n-gram based study, as has been amply illustrated by Biber et al.´s *Longman Grammar* (1999).

2.4.7 Summary and implications for the present study

Subchapter 2.4 has outlined the major methods employed in frequency-based phraseology, as well as the potential questions or issues they may pose. This final section briefly summarises what bearings those methodological issues have on the present research, and what methodological choices they have prompted me to make.

This thesis adopts a corpus-driven rationale, aiming to explore corpus data while understanding frequency and recurrence as the chief and basic indicators of phraseological revelance of patterns. At later

[20] Biber and Conrad (2009) define a register based on its linguistic aspects and the communicative contexts in which these aspects are usually used. "The underlying assumption of the register perspective is that core linguistic features like pronouns and verbs are functional, and, as a result,particular features are commonly used in association with the communicative purposes and situational context of texts" (ibid., p. 2).

stages of the research (case study II and III), the role of collocation strength will also be taken into account, paying attention to making informed choices of particular statistical measures.

Frequency-based phraseology is characterised by a broad range of terminology, specific terms used in a given study typically reflecting the methods employed (Granger & Paquot, 2008). I will be using the term n-gram for automatically retrieved contiguous sequences of n-words or lemmas. N-grams are viewed as a first step of the analysis, aiming to identify frequently recurrent items. Within n-grams I will be identifying patterns, by which I mean meaningful and structurally complete phraseological units.

As regards methods, it has been suggested that much research in corpus phraseology follows one of two basic approaches: either an in-depth study of a small corpus, or a relatively broadly conceived study of a larger corpus (Stubbs, 2002). In this thesis, I will attempt to make use of each of these approaches: the former will be used for pilot probes, the latter for the ensuing case studies. Throughout the three case studies, the n-gram method will be gradually adapted and tweaked through experimenting with length, introducing lemmatisation, empty slots, and positional variation - these parameters have been identified as important in previous research (Čermáková & Chlumská, 2017; Cheng et al., 2006 among others).

Bearing in mind that n-gram extraction alone may tend to be rather coarse-grained, yielding a number of results of "little phraseological interest" (Altenberg 1998, p. 102), n-grams will also be combined with collocation analysis, and with pre-selected lexical items (respectively), in order to utilise all the potential avenues towards revealing the phraseology of corpus texts (Čermáková & Chlumská, 2016; Stubbs, 2002) and to reveal how patterns function in context.

Since recommendations regarding optimum n-gram lengths vary, this question will be given priority and closely explored in the initial case study. In line with the corpus-driven design of this thesis, my approach to n-gram length will be experimental and bottom-up; however, the results will be interpreted with regard to Calc´s (Cvrček, 2021) suggestion that slightly longer n-grams for English than for Czech may ensure better comparability in this language pair.

The following subchapter discussed two more dimensions along which phraseology will be examined: cross-linguistic differences between English and Czech, and the variation of phraseology between different registers. The contrastive outlook of this thesis aims to contribute to filling a gap in contrastive corpus phraseology (Granger, 2014). The typologically distant language pair of Czech and English poses an interesting challenge as well as an opportunity to compare the efficiency of the n-gram method as applied to each respective language. Secondly, phraseological sequences are known to be highly sensitive to register – speakers employ different sequences in each register, reflecting their comunicative needs and the major functions associated with the particular register. Therefore, each of the following case studies will examine a different register and compare its phraseological characteristics between Czech and English.

3 Case study I:

An n-gram-based analysis of Czech and English parliamentary debates. In search of optimum n-gram length

3.1 Introduction

This chapter presents the results of a small-scale exploratory pilot study with a predominantly methodological focus. It aims to describe a selected register through n-gram analysis and to explore its characteristics which will be revealed by a functional analysis of n-grams of various lengths.

I will be analysing parliamentary debates in Czech and English, representing a highly specialised register whose functions are expected to shape its linguistic properties in salient ways. The n-gram method has been chosen because it seems well-suited to characterise a particular register (as discussed in chapter 2).

Register is defined as "a variety associated with a particular situation of use", based on the assumption that "linguistic features tend to occur in a register because they are particularly well suited to the purposes and situational context of the register" (Biber & Conrad, 2009, p. 6).: The functions of recurrent word combinations have been shown to be strongly influenced by register (Biber et al., 2004) or even sub-register (Gries et al., 2011; Groom, 2019). N-grams are a suitable method of pinpointing such recurrent linguistic features in order to characterise a selected register. Frequent word combinations will reflect the pervasive features of the register at hand: in terms of semantics, recurrent themes will be reflected in frequent content/aboutness words contained in n-grams. Word combinations will also pinpoint frequently employed grammatical word combinations, which serve important functions in structuring the text, for example forming means of textual cohesion (Hunston, 2008).

The study focuses on a crucial methodological issue in n-gram-based research: determining the most efficient and informative length of n-grams for examining a given register (Biber & Conrad, 2009; Hyland, 2008). As indicated by earlier contrastive research, the optimum length of n-grams is not only register- but also language-specific (Cvrček & Václavík, 2015; Granger, 2014; Hasselgård, 2017). Applying the n-gram method to corpora of typologically different languages, such as English and Czech, seems particularly "challenging" (Čermáková & Chlumská, 2017, p. 76). As a result, different studies have employed different n-gram sizes. In line with Biber et al.´s (2004) findings, Ebeling and Ebeling (2013) opt for trigrams, while Mahlberg (2013) extracts 5-grams. For Czech, Cvrček and Václavík (2015) recommend 4-grams; however, as pointed out by Hasselgård (2017), for inflectional languages shorter grams typically work better.

In the present study, I explore and compare the results yielded by n-grams of various lengths, starting with bigrams and gradually extending them up to 10 positions. The ultimate aim is to determine the most informative n-gram lengths for examining the selected register. If the optimum lengths for the respective languages differ, I presume this to be related to typological differences. Another objective of this study is to provide an n-gram based description of the register of parliamentary debates in Czech and English.

3.2 Material and method

The study is corpus-driven: it proceeds from the identification of 2-10-grams (i.e. continuous "recurring strings, with or without linguistic integrity" (Lindquist & Levin, 2008, p. 144)) to the qualitative functional description of n-grams of various lengths, and identification of meaningful and structured patterns (ibid.) within those n-gram strings.

Two small corpora were custom-compiled for this study. The material was obtained from official records of British and Czech parliamentary debates published online. For English I have drawn on the Hansard (*House* *of Commons Official Report*), using the records from two randomly selected days (16 June 2015 and 11 January 2017; total 10 698 types, 243 366 tokens).[21]

For Czech, the transcripts of parliamentary proceedings published on the official website of the Chamber of Deputies (Poslanecká sněmovna) proved inconvenient due to their formatting, which does not allow for a bulk download of the whole day´s proceedings – the transcripts are structured as an intertext list linked to individual short sections. This makes data retrieval extremely labour-intensive. Therefore I have only used the transcript of 1 whole day obtained from the Chamber of Deputies´ website[22] (25 November 2013; 7842 types, 43 588 tokens). To complement this sample, I have drawn on an existing corpus of parliamentary proceedings, CzechParl (Pražák & Šmídl, 2012) (137 531 types, 241 357 tokens). This combination of two sources ensured better comparability with the English corpus: while the Hansard material consisted of full transcripts of each day, the CzechParl corpus is composed of excerpts, which may result in some parts of the proceedings being omitted; the inclusion of the web transcripts at least partially controls for this.

The data was not lemmatised. A brief quantitative look at the two corpora reveals a difference in the type-token ratios between the two languages (cf. table 1). For the English corpus the TTR was 0.04, while for Czech it amounted to 0.51. This probably reflects Czech being considerably more morphologically variable than English – this aspect will be taken into account in case study II, employing lemmatised data in an attempt to overcome this cross-linguistic difference.

[21] https://hansard.parliament.uk/pdf/commons/2015-06-16
[22] https://www.psp.cz/eknih/2017ps/stenprot/index.htm

	English	Czech
types total	10,698	145,373
tokens total	243,366	284,945
TTR	0.04	0.51

Table 1. Type-token ratios of the two corpora examined

The two corpora were subsequently loaded into AntGram (Anthony, 2017). For each corpus, n-grams were extracted, n ranging between 2 and 10, the minimum frequency was set at 20 per million words, punctuation was excluded. For each length a separate n-gram search was performed in the respective corpora, resulting in 18 n-gram frequency lists. An arbitrary cut-off rank of 80 was established in order to ensure easier manageability. The n-gram lists were then compared for each language separately and analysed qualitatively. First, I identified overlapping n-grams and marked these up. For example, the 3-grams *my hon friend + hon friend the + friend the member + the member for* in fact form a longer sequence together: *my hon friend the member for.* Then, I annotated the n-grams according to their grammatical structure and, if applicable, to discourse functions. These two steps are aimed towards the identification of patterns: formally coherent sequences assocaited with a particular meaning and/or function. Finally, I tried to identify the optimum n-gram length for examining each language, considering the frequency and the amount of register-specific information (structural and functional) revealed by n-grams of each length.

As will have become apparent, the frequent overlaps between n-grams pose a major methodological issue. Firstly, there were numerous overlaps between the sequences identified within each n-gram length, as in the abovementioned 3-grams *my hon friend + hon friend the + friend the member + the member for*. Secondly, shorter n-grams were very often contained within longer ones, as in *hon friend/hon friend the/hon friend the member*. Hence I have opted for a purely qualitative approach and will not attempt to give an overview of the total numbers of n-grams identified within each length. I believe that due to the extensive degree of overlapping, such total counts in themselves are not informative. Instead, a qualitative analysis of the data, focussing on any recurrent meanings and functions conveyed by the phraseological units, is expected to be more suitable and efficient for characterising the phraseology of parliamentary debates through recurrent functional units – patterns.

3.3 Results

3.3.1 Register characteristics

The material displays the specific features of the parliament register, allowing for a functional classification of some register-specific n-grams. Broadly speaking, similar discourse functions were identified in both languages; however, there are qualitative differences, pointing to potential different cultural norms and practices shaping the parliament interactions in the respective languages.

The nature of the register is reflected in some patterns fulfilling functions determined by the particular situational characteristics (Biber & Conrad, 2009, p. 31) of parliament register, namely the discourse participants, their relations and a particular established manner of communication they adhere to. For instance, in the English corpus the social roles of the participants are reflected in highly stylistically marked honorifics (*My hon. Friend the Member for*) and other politeness markers. The Czech debates were characterised by frequently recurrent, formally fixed, register-specific performative formulae (*zahajuji hlasování ptám se kdo je pro kdo je proti – "I commence the vote who is for who is against"*) related to the specific communicative purposes of the parliament register. Patterns containing content words or deictic elements (cf. table 2), found in both languages, relate to the subject matter of the discussion.

pattern group	CZ	EN
performative formulae	performatives *zeptám se zda* instructions *slova se ujme pan*	performatives *I beg to move* instructions *will the hon gentleman give way*
honorifics, politeness markers	address terms *vážené paní poslankyně, vážení páni poslanci* thanks *děkuji vám za pozornost; děkuji za slovo*	address terms (in/direct) *my hon friend is absolutely right;* *it is a pleasure to serve under your chairmanship, Mr X* thanks etc. *I congratulate; I thank*
content words	legal terms *ve znění předchozích předpisů* figures, dates *v roce dva tisíce*	legal terms *secretary of state;* *EU referendum bill*

deictics	*to je; a to; k tomu;*
	v současné době;
	proti tomuto návrhu
		that is why; that is what

modal	epistemic – hedging *(já) si myslím (že);* *podle mého názoru*	epistemic *I think that; I am sure* deontic *I want to; we need to; it is important*
fixed grammatical patterns (no content word)	N/A	existentials *(there is…);* *this is not just about*

Table 2. Functional classification of n-grams in both languages

Overall, the findings suggest that the criteria of typological characteristics and register are closely interrelated in n-gram-based analysis. A detailed discussion of findings for each language follows.

3.3.2 English

The English data were characterised by patterns containing a content word, which mostly fall under the heading of legal/political terminology, or the names of acts and bills discussed during the particular session, as shown by exx. 1-3.

1) the government, the electoral commission, secretary of state
2) european union referendum bill 16 june 2015
3) the political parties elections and referendums act

Another group were patterns containing politeness markers. The politeness is usually based in the MP´s status within the Parliament and expressed through honorifics. Honorifics whose usage is limited to the parliamentary setting are employed: *my honourable friend, the honourable member/gentleman* etc., used among MPs, and *your chairmanship,* used towards the Speaker. They provide an insight into the practices of interaction among Members of Parliament. Traditionally, British MPs only interact through the Speaker, not addressing each other directly; as a result, even members who are present are referred to in the third person. Example (6) is a fixed honorific phrase used to address the Speaker.

4) my hon friend the member for
5) the hon member for leeds
6) it is a pleasure to serve under your chairmanship

Closely linked to honorifics are performative statements (ex. 7) and performative formulae, serving to give instructions and organise the session (ex. 8), even though this latter group was smaller and less varied than its Czech counterpart (discussed in the following section). Such statements are uttered by an MP and addressed either directly to the Speaker, or to another MP via the Speaker, as in 8.

7) I beg to move
8) will the hon gentleman give way

This raised the question of the Speaker´s turns, since these did not occur in the n-grams. A closer inspection of the source text revealed that the Speaker´s utterances are in fact fairly sparse and mostly limited to one word (demanding "*order*"), which makes them unidentifiable through an n-gram search. Still, "*order*" could be added to the performative formulae category.

Further, modal patterns were found, expressing either epistemic modality, some of which may function as hedges (ex. 9, cf. Czech counterparts with the same function); or deontic modal meanings connected to evaluation (10).

9) I think that, I am sure
10) I want to, we need to, it is important

Finally, a structurally defined category of patterns comprised fixed grammatical sequences which did not contain a content word, but rather served a particular grammatical (existential constructions, ex. 11) or textual function (such as emphasis, ex. 12).

11) there is ___
12) this is not just about

3.3.3 Czech

Similarly to English, the Czech corpus contained examples of legal terminology (exx. 13-15), as well as a related content pattern category, which was however not represented in the English data – n-grams containing various figures and dates (ex. 15).

13) ve znění předchozích předpisů
"in the words of previous regulations"

14) vládní návrh zákona kterým se mění zákon číslo
"the government bill which changes act no."

15) v roce dva tisíce
"in the year 2000"

Like their English counterparts, Czech MPs employ politeness markers/honorifics (exx. 16-17). The politeness is again rooted in their MP status; nevertheless, the Czech terms of address prove less register-specific than the English ones: they essentially comprise the adjective *vážený* and a job title, cf. exx. 16-17. Analogous terms of address are used in standard Czech in formal written communication,[23] which seems in line with the formality of the parliament register. Notably, in contrast to their British counterparts, the Czech MP's interact with each other directly.

16) vážená paní předsedající, vážené paní poslankyně, vážení páni poslanci
"dear Mrs Chair, dear ladies-MPs, dear gentlemen-MPs"

17) pane ministře
"Mr Minister"

The use of epistemic modal hedges as in exx. 18-19 may also be regarded as a politeness marker.

18) (já) si myslím (že)
"I think (that)"

19) podle mého názoru
"in my opinion"

One type of n-grams abundant in Czech is represented by performative formulae, whereby the speaker performs a particular action within the parliamentary session. Some of them contain a performative use of communication verbs (in bold in exx. 22-23). Notably, a closer inspection

[23] https://prirucka.ujc.cas.cz/?id=850&dotaz=osloven%C3%AD

in context reveals that when functioning as instructions, these formulae are normally addressed from the Speaker towards the MPs, reflecting the Speaker´s status of an authority figure.

20) ať stiskne tlačítko a zvedne ruku kdo je proti tomuto návrhu
"let push the button and raise their hand whoever is against this bill/proposal"

21) slova se ujme pan
"the word goes to Mr"

22) ptám se kdo je pro kdo je proti
"I ask who is for who is against"

23) zahajuji hlasování
"I commence the vote"

Bigrams are usually not employed in semantic analysis, since they tend to consist predominantly of function words, which typically rank among the most frequently recurring combinations. However, although the bigrams identified in the Czech corpus did comprise mostly function words, arguably they still reveal a remarkable aspect of the register at hand. Demonstrative pronouns and other deictic words were particularly frequent in bigrams, which indicates explicit ties to the immediate co(n)text of the utterances (ex. 24). Looking beyond bigrams, longer deictic patterns were also identified within 3-grams, confirming this tendency (25-26). The English data likewise contained some deictic n-grams linking two utterances (27).

24) bigrams: a to, to je, k tomu, na to, o tom, to znamená
"and it", "it is", "to it", "on it", "about it", "this means"

25) v současné době
"currently"

26) proti tomuto návrhu

"against this bill"

27) that is why; that is what

This suggests that deixis may have a particular functional load in parliamentary debates, given the frequent reference to the subject of the debate, and the need for precision and clarity. The previous findings about address terms support this, as address terms are considered a subtype of situational deixis (Hirschová, 2017). Admittedly, this suggestion is only tentative as the corpus employed here was fairly small; this potential tendency would need to be verified against a reference corpus.[24]

The longest grams extracted in my study (n = 9 or 10), although few, proved interesting; they often comprised clausal structures. These long sequences manifest a high degree of overlaps, indicated by the relatively low type-token ratio of 0,18. Upon closer inspection many of these overlaps were collapsed, highlighting the importance of follow-up qualitative analysis on this approach.[25]

In ex. 28, the n-grams were subsumed under one long sequence, even though the component sequences did not always overlap exactly. However, this points to an underlying pattern allowing for a limited degree of syntactic variability, as indicated by the parts in brackets.

28) zahajuji hlasování (a) (ptám se) kdo je pro (a) kdo je proti v zasedání (pořadové) číslo X
"I commence the vote (and) ask who is for (and) who is against in session (ordinal) number X"

In line with this finding, some of the shorter n-grams recurred with

[24] A quick keyword analysis (of the Chamber of Deputies transcript analysed in this study) through K-Words with the general SYN2015 as a reference corpus did not identify any demonstrative pronouns as keywords.

[25] As stated earlier, these overlaps make it very problematic to state precise counts of the individual n-gram lengths.

limited lexical variability, pointing towards patterns with an open slot:

29) zákon číslo X/Y
"act no. X/Y"

3.3.4 N-gram length: a contrastive overview

Earlier studies (e.g. Čermáková & Chlumská, 2017; Hasselgård, 2017) indicated that the length of n-grams relevant for a particular study is predominantly determined by two factors: the nature of the register in question, and the typological characteristics of the language. The factor of typological features has likewise proven relevant in the present study.

The structure of the English bigrams reveals the analytic character of the language: they comprise combinations of function words – mostly articles and prepositions (*of/in/to/on/for the*, *that/and the*, *it is*, *I am*). Lexical words are few in English bigrams, mostly represented by honorific terms (*hon friend*, *member for*) or content expressions (*the government, european union*).

Content words are more widely represented in 3-grams and longer sequences. 3-grams contain lexical aboutness words; there are also fixed constructions, mainly existentials. 4-, 5-, and 6-grams also comprise aboutness words, as well as frequent fixed phrases, such as honorifics.

In English, long n-grams of n=7 or more were not very informative due to their low frequency of occurrence both in terms of tokens (the highest frequency was 50, followed by a steep drop) and types – there were overlaps, some long n-grams were even instances of repetition of a shorter sequence, either caused by the speaker repeating a phrase, or possibly by a mistake in the transcript. The few informative[26] long grams

[26] Relatively few of the very long n-grams lend themself to semantic or functional characterisation since a number of them are in fact instances of repetition of shorter n-grams. This may be due to a flaw in the text file or perhaps the inclusion of captions (cf.

were law/bill titles, and highly register-specific honorifics, which had been represented among shorter sequences as well (*my hon friend the member for south norfolk mr bacon; it is a pleasure to serve under your chairmanship mr*).

In Czech, more lexical words are represented in bigrams, comprising forms of address (*vážená paní, pane ministře*); discourse-organizing and stance bigrams (*děkuji za* 'I-thank for', *myslím že* 'I-think that'); and aboutness words (e.g. legal terms, figures). Apart from some lexical word bigrams, the function words comprise deictic expressions. As suggested earlier, these possibly reflect the speaker´s preoccupation with precise reference.

Still, even in Czech, bigrams do not seem to allow a sufficiently thorough insight into the nature of the parliament register. Moving on to 3-grams, there is a vast degree of overlapping – several 3-grams forming longer sequences (cf. *prosím pane poslanče + pane poslanče máte + poslanče máte slovo + máte slovo děkuji*). The same is true of 4-grams. Interestingly, overlapping 3- and 4-grams tend to form 5-grams, but not longer patterns. In other words, while 5-grams often comprise chunks which were also retrieved among shorter grams, longer sequences (roughly 7-grams and longer) do not considerably overlap with shorter ones; typically they are also less frequent.

Unlike English, long n-grams in the Czech corpus proved revealing, albeit relatively few as well. They contained performative formulae limited to the parliament register. These were more or less fixed, although they did allow for some limited lexical variability in the form of optional elements (ex. 28 above). The following section suggests that such formulae reflect the rigid norm governing the organisation of parliamentary debates.

european union referendum bill european union referendum bill 16 june).

In summary, while n-gram based research often focuses on n-grams of up to 5 words, the results suggest that the parliament register relies on relatively longer n-grams (6-8 words in Czech, 5-6 in English), which warrant examination. These patterns correspond to relatively fixed phrases performing a highly specific function in the communicative situation, whose usage appears limited to this particular register.

3.3.5 Fixed sequences as a reflection of register norms

As the register of parliamentary debates is elaborated and specialised to a large extent, this is reflected in equally specialised word combinations, allowing for the establishment of relatively long stable sequences (cf. ex. 28 above). Any register is governed by norms emerging from social expectations and needs shared by its users/participants (Dovalil, 2012, p. 139 with reference to Luhman 2008: 31), which influence their linguistic production. The norms are derived from the recurrent linguistic behaviour within the register, i.e. they are essentially frequency-based. In turn, recurrent instances of language use as per this norm further reinforce the norm. Thus, norms are constantly re/created and stabilised throughout interactions; and there is a bilateral relationship between norm and usage (cf. Verhagen, 2015).

The parliament register is evidently shaped by the need for stability, precision, reliability of the statements and the speaker´s commitment to their truth value. The community of speakers is restricted and they share the awareness of all their communication being recorded, stored and made publicly available. Therefore, the speaker is expected – and likely – to conform to their audience´s expectations and follow the norms associated with the debates.

The frequently employed patterns are to some extent flexible (allowing for "open slots", i.e. limited lexical variability), but creative usage is not favoured. This is likely due to the need for precision (cf. correct use of address titles, referring to legal concepts, laws, bills etc.), as well as for clarity and legitimacy within the register (e.g. fixed performative formulae, which give rise to no doubt about the speaker´s communicative intention).

3.4 Conclusion

Broadly speaking, the optimum length of n-grams appears to be different in Czech and in English. While the differences can be accounted for by the typological differences between the languages (while English is predominantly analytic, Czech is a synthetic language with rich inflection), some specific features of English and Czech parliamentary debates can also be pointed out, e.g. the highly register-specific honorifics found only in English.

This pilot study has confirmed the potential of n-grams to efficiently reveal cross-linguistic similarities as well as parallels, and to pinpoint the typical, distinctive phraseological features of a highly specialised register, even when using a limited amount of material.[27] The aim was to identify the most suitable n-gram length for the purpose of characterising the phraseology of the parliament register. However, the results suggest that the question of "which n-gram length is best?" should perhaps be reframed as "what can each n-gram length reveal about the register at hand?".

Essentially, the results of the study indicate two possible approaches. One is comparing various n-gram lengths, each of which will reveal patterning of different sort, and combining these complementary probes to allow a comprehensive insight into the phraseology of the register at hand. The advantage of this approach is that it provides a varied and detailed overview of meanings and functions conveyed by phraseological sequences. Longer n-grams may point out extremely rigid word combinations which are likely to be highly functionally specialised, and whose use may be limited to the particular register. Short sequences tend

[27] Cf. Gries et al., 2011, who suggested a similar conclusion: a relatively small slice of n-gram data can pinpoint the salient features of a register with surprising efficiency.

to contain a large proportion of grammatical words, which may reveal some interesting patterning nonetheless, as was the case with deixis in the present study. The longer the n-gram, the less frequent it will be; consequently, the findings obtained through different n-gram sizes would need to be normalised. Another caveat to be borne in mind is the well-documented register effect on phraseological units (Biber et al., 2004): a given n-gram length may point to very different patterns in different registers. Hence, employing a variety of n-gram lengths should be associated with a corpus-driven, exploratory approach.

An alternative option is to adhere to a selected mid-length n-gram size(s). This size should ideally return a sufficient number of results to provide a quantitatively valid picture of register phraseology (i.e. avoiding very long and thus infrequent sequences); on the other hand it should not retrieve too many n-grams (i.e. avoiding bigrams), thus ensuring manageability and enabling a complementary qualitative examination. The results suggest that 4- to 5- n-grams are a reasonable compromise for this purpose.

As shown in this case study, major methodological issues were linked to overlapping n-grams. The overlaps identified in the parliament corpora are or two sorts: either shorter sequences are contained in longer ones, i.e. they are represented multiple times in the results, if comparing various n-gram sizes; or several n-grams are composed of the same words/lemmata, only differing in their order. At any rate, these overlaps should be taken into consideration while classifying and evaluating the results. In the present study I did not attempt to provide a precise quantitative description of the n-grams retrieved, since accomplishing it would have been complicated and imperfect, and given the primarily qualitative focus of the study, this aspect was deemed expendable. The problem of n-gram overlaps also speaks in favour of focussing on the identification of patterns – recurrent structurally complete units conveying a particular function (Hunston & Francis, 2000), which may comprise several overlapping n-grams and may even allow for some degree of variability (e.g. open or variable slots).

A question which yet remains to be explored is lemmatisation: what dis/advantages are brought about by using lemmatised data (or not), and possibly also what degree of lemmatization is favourable (cf. potential drawbacks noted by Čermáková & Chlumská, 2016, as discussed in chapter 2).

In the following chapter, the n-gram method will be tested on a larger corpus, which however represents a type of register not as narrowly functionally specialised as parliamentary debates. This will make it possible to modify the method and further test its efficiency.

4 Case study II: Prepositional patterns in English and Czech newspaper reporting

4.1 Introduction, hypotheses

The aim of this study is to explore the phraseological features of the newspaper register, comparing these between English and Czech, and focussing on means of discourse organisation and structuring the text.

This study is conducted through a combination of several corpus-driven methods: an initial exploratory n-gram search is conducted in order to reveal the types of phraseological units employed in the data and suggest an inductively conceived classification of these units, based on their semantic properties and textual functions. Next, a selected group, prepositional patterns, is analysed in detail. The n-gram extraction is complemented by a collocation analysis. The aim is to reveal how the recurrent phraseological sequences are employed in context. Prepositional patterns are expected to participate in building textual relations (Hunston, 2008) as well as to manifest particular semantic prosodies (Partington, 2004).

4.2 Theoretical background

4.2.1 Newspaper register

In this chapter I focus on the phraseology of newspaper language. Within the larger scope of the dissertation, newspaper language represents a somewhat less narrowly specialised register than that of political debates, which were analysed in the previous chapter. Like children´s fiction (examined in the following chapter), newspapers are a written register. Both fiction and newspapers are widely accessible to most language users, but they are distinguished by their focus and communicative function: "aesthetic and recreational" as opposed to “informational”, respectively (Biber et al., 1999, p. 25).

Using corpus methods and focussing on English newspaper data, Silvennoinen (2017) identified systematic quantitative as well as qualitative differences in the employment of grammatical patterns (expressing contrastive negation) both between different registers (newspapers and conversation) and subregisters (different types of newspaper texts). In the present study I will not be analysing newspaper texts along the subregister dimension, i.e. the aim is not to systematically compare different thematic sections of newspapers, or different types of newspapers (e.g. so-called quality papers as opposed to tabloids). Instead, bearing in mind the larger scope of this dissertation project, I approach the dataset at hand as a relatively homogeneous representation of the register of newspaper texts.

The newspaper register is potentially interesting due to its functional properties. Its primary focus is indeed presenting information, reporting on current events etc. However, its function goes beyond a purely informational one: it also offers a particular interpretation of the reality, since “newspaper editors make selections from what they could report, when deciding what they will report” (Scott & Tribble, 2006, p. 162). Based on this aspect of the newspaper register, I assume that to obtain comprehensive characteristics of newspaper discourse, it should be examined from the perspective of semantic/evaluative prosody and semantic preference (e.g. Partington, 2004; Sinclair, 2004), i.e. specifically looking at the tendency of patterns to be involved in conveying positive or negative evaluative meanings (semantic prosody), and/or to co-occur with items from a particular semantic field (semantic preference).

In the previous chapter, the n-gram method was applied to political debate transcripts, the results suggesting that this approach can effectively reveal salient features limited to that highly specialised register. In the present chapter the same n-gram method will be employed to examine a less specialised register, to test whether it may be similarly effective in revealing its phraseological properties.

To summarise, the aims of this study are:

- to identify and describe the functional types of phraseological units employed in newspaper texts,
- to provide a detailed analysis of a selected pattern group – prepositional patterns,
- to reveal how prepositional phraseologies are involved in construing textual relations and evaluative meanings in newspaper texts. To this end, n-grams will be complemented by collocational analysis.

4.2.2 Grammatical words in patterns

Using closed-class items such as prepositions as the starting point towards analysing the semantic features of a register may seem counter-intuitive, since these words are traditionally perceived as having grammatical functions only. However, evidence from corpus phraseology suggests that meanings are conveyed primarily through word combinations rather than isolated words (Ebeling & Ebeling, 2013; Groom, 2010, p. 61). Consequently, "the supposedly meaningless closed-class words make just as important a contribution to the overall meaning of each phraseology as do the open-class items [...], and therefore constitute equally valid starting points for a semantically-oriented analysis." (Groom, 2010, p. 62)

The present chapter is based around the presumption that grammatical words may in fact be a highly suitable basis for characterising a particular register, because they are highly frequent and evenly dispersed throughout discourse (cf. Sinclair 1991), occurring in a range of cotexts and uses.

Given that closed-class words are the commonest words in virtually all corpora, it follows that an analysis based on even a small selection of them will account for a far greater proportion of the data as a whole than can be achieved through an analysis of even a large selection of open-class items (cf. Zipf 1935; Sinclair 1991, 1999). (Groom, 2010, p. 71)

Therefore, closed-class words, used as the starting point of a phraseological analysis, may even help identify "a much wider range of phraseological phenomena than might otherwise be possible", and point towards phraseological units of varying degrees of formulaicity (Groom, 2010, p. 71).

The relevance of grammatical words is further highlighted by their involvement in discourse structuring, fulfilling a variety of textual functions and contributing to textual coherence; hence on a larger scale these grammatical patterns can also help reveal pervasive discourse patterning (Hunston, 2008).

4.3 Data & method overview

First, a pilot n-gram-based probe into the newspaper register was conducted on the Bank of English™ corpus. This pilot identified some methodologically problematic areas – the findings were taken into account when designing the next, contrastive study, which forms the main part of this chapter. The contrastive study then employs a combined approach: n-grams containing prepositions are selected, and patterns containing selected prepositions frequent in newspapers are analysed qualitatively. I examine prepositional phraseologies, their involvement in structuring newspaper texts and their semantic prosody and preference.

4.4 Pilot study: exploratory n-gram analysis of English newspapers

4.4.1 Aims

This initial probe aims to test whether simple n-gram extraction may provide insight into which phraseological units frequently recur in newspaper corpora and what textual functions they fulfil. A similar approach has proven efficient in the previous study on parliamentary debates. In that study, n-grams helped me identify recurrent patterns which included highly register-specific, formally fixed phrases. This has raised the question of whether n-gram analysis alone could yield equally informative results if applied to a register less specialised than the parliamentary one. The aim of the present chapter is therefore to test the n-gram extraction method on newspaper data in order to reveal to what extent it can highlight prominent features of newspaper discourse.

4.4.2 Data & method

The data employed in this pilot study come from the 553 million token Bank of English™ (henceforth BoE) corpus, a representative sample of the 4.5 billion word COBUILD corpus, which comprises a variety of registers including newspapers, fiction, spoken language and magazines, representing 8 different regional varieties of English. Newspaper texts form roughly a half of the corpus.[28]

For the purposes of this pilot probe, a subcorpus was created within BoE, limited to UK newspapers published between 1995 – 2005 (125 million tokens; the majority published between 2001 – 2005). This temporal restriction was applied in order to make the resulting subcorpus potentially comparable with the Czech SYN2009PUB (cf. subchapter 5), even if smaller.

Next, 4-grams were extracted from this UK newspaper subcorpus. A cut-off point (admittedly arbitrary – motivated by practical considerations) was placed below the top frequent 300 n-grams. Within these n-grams, patterns were identified, examined in their contexts and categorised into

[28] For more details see https://wordbanks.harpercollins.co.uk/Docs/WBO/WordBanksOnline_English.html

groups in a bottom-up, inductive manner, following a combination of criteria:

- formal (does the pattern contain a particular part of speech?),
- semantic (does a pattern refer to a semantically defined group of referents? can the pattern be assigned a particular semantic role?),
- functional (does the pattern fulfil a particular textual function?).

The resulting pattern groups are described in the following section.

4.4.3 Results

Table 3 summarises the pattern groups identified on the basis of formal, semantic or functional features observed in the n-grams.

pattern group	example	raw frequency - pattern types
function words	. "It 's : "I ´m	101
speaking verbs	. He said: He added: "	56
content words	Football; Match report ; World Cup qualifier the World Trade Centre	45
time	for the first time of the season.	23
space	the top of the in the world.	18
partitive	is/was one of the one of the best one of the most	11
thinking verbs	I do n't think	8
wanting verbs	I do n't want	6
text organisers	when it comes to	6
time or space	at the end of	5
other (not classified)		21
total		300

Table 3. Pattern groups in BoE newspapers

Before introducing the resulting pattern groups, I will make two methodological remarks. Firstly, a look at the results reveals numerous overlaps between the n-grams identified, as illustrated for instance by these three n-grams, which ranked consecutively on the frequency list.

- ; Football ; World freq 467
- Football ; World Cup freq 422
- ; World Cup qualifier freq 377

This confirms that a more sophisticated method would be needed to collapse such overlapping n-grams together. However, for the purposes of identifying patterns within n-grams and categorising them into functional groups, even this rough frequency list will presumably suffice.

Secondly, as shown predominantly by the function word patterns and speaking verb patterns, punctuation has a strong presence in the patterns, highlighting the importance of patterning around syntactic boundaries or related to introducing direct speech. This suggests that including punctuation in the n-gram search may be desirable; but the adjacent cotext will need to be examined to reveal more about the functions of such patterns.

Next I discuss the largest pattern groups identified.

4.4.3.1 Function words

The most widely represented group of patterns included function words, namely a personal pronoun (most often *it* or *I*), followed by a verb, usually the copular *be*. Typically, these patterns were sentence-initial, as indicated by the preceding punctuation.

4.4.3.2 Verbs of speaking

Speaking verb patterns are in fact closely linked to the previous group of function word patterns. The verbs of speaking such as *say, add* serve to introduce a direct quotation, which in turn often opens with a function word sequence. This is confirmed by the occurrence of patterns such as exx. 1 and 2.

1) said: "It
2) added: "I

Additionally, the speaking verb n-grams point towards an imbalance in the representation of male and female speakers. In patterns introducing direct speech, the personal pronoun *he* is considerably more frequent than *she*, indicating that male speakers are more often cited in this newspaper collection. The two most frequent n-grams of the whole set are the following overlapping n-grams (exx. 3 and 4):

3) . He said : (freq 4,845)
4) He said : " (freq 4,583)

Further down the frequency list, these two n-grams are followed by variations on *he added, said, says*, as well as *spokesman said*, combined with punctuation in various constellations. Within the top 300 n-gram type sample examined, a total of 4 n-gram types (disregarding punctuation)[29] and 6,714 n-gram tokens referred to a male speaker. By contrast, n-grams in which female speakers were represented (containing *she said/says*) comprised only 2 n-gram types (again disregarding punctuation) and 1,794 n-gram tokens.

This clear indication of women speakers being underrepresented may be interpreted in light of the findings from the content word pattern class, which indicates that the topic of sports receives major coverage in the dataset at hand. Sport disciplines referred to in the n-grams included

[29] Strictly speaking, a total of 15 such n-gram types were identified; yet there was a large degree of overlaps and a number of patterns were only differentiated by punctuation, cf. exx. 3 and 4. Hence punctuation was disregarded in this particular count.
For a detailed discussion of the concept of *n-gram types X n-gram tokens,* see subsection 5.2.

football and rugby, both of which are (or were at the time of the publication of these newspaper texts) arguably predominantly the domain of sportsmen. This is illustrated by two other patterns found in the data, occurring in headlines or possibly picture headings (exx. 5, 6).

5) MAN OF THE MATCH (freq 302)
6) DREAM TEAM STAR MAN (freq 199)

Together with the prevalence of sports-related patterns (discussed in more detail below) this may explain the tendency for male speakers to be quoted more frequently than female ones in this dataset.

However, it should be noted that the inclusion of punctuation in the n-gram search may also play into the prominence of these quoting patterns. Due to including punctuation, there were numerous overlaps between the n-grams retrieved (cf. footnote 29). As a result, a number of the *he said* patterns may well be counted multiple times, which increases the total count. Such overrepresentation of patterns due to overlaps has proved to be a major drawback of the n-gram method in case study I, where it effectively prevented a precise quantification of the patterns. In case of the quoting patterns in newspapers, such overrepresentation may be even greater because of the immense frequency of punctuation.

4.4.3.3 Content words and proper nouns

This group of n-grams contained a content word – denoting various entities, events, or topics referred to in the newspaper texts, reflecting the "aboutness" of the texts (Bondi, 2010, p. 7). The vast majority of content word patterns (36 out of 45 pattern types) referred to sports, chiefly football, as in exx. 7, 8.

7) Football ; Match report
8) ; World Cup qualifier

The remaining 9 pattern types included e.g. the title of a newspaper contained in the corpus (ex. 9) or references to current affairs, specifically events related to 9/11 (exx. 10, 11). In exx. 9 and 11 the frequency of the pattern is also due to it being a proper noun, recurring invariably in the same form.

9) News of the World
10) ; War on terror
11) the World Trade Center

The content word patterns highlight the prevalence of certain topics which receive a large amount of media coverage. However, the overwhelming presence of patterns referring to sports seemed peculiar; intuitively, it seems unlikely that the topic of sports should occupy such a large portion of newspaper discourse. Therefore, another n-gram search was performed, this time with the data limited to broadsheet newspapers only, reducing the corpus size from the original 125 million to 73 million tokens (cf. table 4).[30]

subcorpus	BoE subcorpora	newspapers included	tokens
tabloids	sunnow	Sun, News of the World	51,805,654
broadsheets	times	Times, Sunday Times	46,759,194
	brregnews	Regional Newspapers	21,029,439
	brnews	The Independent	6,006,167
total			125,600,454
total without tabloids			73,794,800

Table 4. Corpus structure

[30] My thanks go to Dr Nicholas Groom for suggesting this step.

The resulting 4-grams were classified following the previously identified pattern groups. The results confirmed the suspicion: the content pattern group in the broadsheet-only data proved conspicuously different than in the broadly conceived newspaper collection (i.e. including both broadsheets and tabloids). Broadsheet content patterns include words referring to more varied referents than the tabloid patterns. Although references to sports were also represented in the broadsheets (e.g. *the World Cup finals*), they were by no means as prominent as previously. Content patterns included proper nouns and referred to a variety of entities likely to be involved in topical issues, namely political (*the House of Lords, the House of Commons*) and other institutions (*Bank of England, Church of England, World Trade Center*), or the titles of public figures (*the Prince of Wales, Secretary of State for*).

In summary, the n-gram extraction has efficiently revealed salient features of newspaper discourse: It has pointed towards the importance of direct speech, and indicated conspicuous aboutness words, suggesting major topics represented in newspaper discourse. It suggested that a focus on grammatical words in the patterns may be revealing, in this case pointing towards an imbalance in the representation of male and female speakers quoted.

The major point of insufficiency in this study, as revealed by a comparison between content n-grams extracted from all newspapers as opposed to broadsheets only, can be described as “the influence of corpus design on the identification of important lexical phrases” (Gray & Biber, 2015, p. 137). The categories of newspapers included in the dataset had a pronounced effect on the semantic nature of frequent n-grams retrieved; this suggests that the factor of register plays a major role and needs to be taken into consideration when interpreting the data.

The problem of the overrepresentation of sports in the n-grams also raises an interesting methodological question: the prevalence of sports-related patterns may be simply due to the language of sports reporting being highly repetitive. Arguably, the sports sections of newspapers are largely focussed on relatively few sports which receive the majority of regular media coverage, such as football. Furthermore, sports are regulated by strict rules and described through established terminology. This may cause sports to be overrepresented as a result of n-grams being employed. In other words, sports may not be more frequently reported in the newspapers, but rather they seem more prominent when the newspaper data are viewed through n-grams.

4.4.4 Implications for further research

This pilot probe has suggested that a simple exploratory n-gram based approach to a corpus of newspaper discourse does not quite suffice to yield satisfactory results in characterising its phraseological properties, although it can serve as a useful first step.

First, a major register effect was noted – patterns associated with topics frequently discussed in tabloid press become prominent in the n-grams, consequently skewing the results. These findings pointed towards a need to restrict the newspaper data to broadsheets, in order to avoid a major bias towards topics such as sports.

For this reason, as well as in order to employ a larger dataset directly comparable with the Czech SYN-PUB corpus, the following case studies will draw on a different source of newspaper data, namely the SiBol corpus (Duguid et al. 2005) for English and SYN2009PUB (Křen, 2009; Křen et al. 2010) for Czech (discussed in more detail below in the respective sections).

Second, the results of the exploratory probe yielded several interesting isolated findings, but broadly speaking they were not very informative with regard to how the recurrent phraseological sequences actually operate in context. The n-gram search identified fragmentary units, whose immediate textual environment would need to be examined more closely in order to reveal how these units contribute to building textual structure. This is further complicated by the fact that a number of patterns overlapped, effectively forming longer sequences; or some pattern groups partially co-occurring (function word patterns and speaking verb patterns introducing direct speech).

In the light of these pilot study findings, as well as taking into account the theoretical assumptions presented in section 2, in the following cross-linguistic study n-grams will be employed as a first step, complemented by an analysis of their collocations and their involvement in conveying evaluative meanings.

Furthermore, to obtain more detailed and comprehensive results, the analysis will be limited to patterns containing prepositions. Prepositions seem a valid starting point for several reasons, some of which have been outlined in section 2 – let us briefly recapitulate them here. Firstly, prepositions are frequent and evenly dispersed in texts; therefore prepositional patterns are an efficient tool to provide a comprehensive portrait of the phraseological characteristics of a corpus, and are able to identify a variety of pattern types fulfilling different textual functions (cf. Groom 2010); they also help reveal how texts are structured (Hunston, 2008). Secondly, prepositions are an interesting – even if potentially challenging – basis for cross-linguistic comparison, due to the overall lack of correspondence between preposition translation equivalents (Klégr & Malá, 2009).

Further, using closed-class words as a stepping stone towards pattern identification helps reveal "repeated sequences of semantic elements which may have a very heterogeneous surface realisation" (Groom, 2010, p. 72), i.e. semantic sequences (Hunston 2006). This in turn allows for a comprehensive analysis of how patterns operate in texts, which textual functions they fulfil and what meanings they convey.

In order to obtain a comprehensive view of the roles played by prepositional patterns in newspaper texts, the following procedure is adopted in the following study. For each language, I first identify 3-, 4- and 5-grams containing any preposition in any slot. These n-grams point towards frequent prepositional patterns occurring in the newspaper register in each language. The prepositional patterns are compared crosslinguistically and two frequent prepositions contained in them are selected for further analysis in each language: *v, na* for Czech; and *of, in* for English. *V* and *in* are chosen also because apart from ranking among the most frequent prepositions, they are translation equivalents, allowing for an additional cross-linguistic comparison, which is relevant to this study.

These selected prepositional patterns are then analysed further: they are classified inductively into groups, based on a combination of criteria, namely their form, semantics, and textual functions. Next, I search for the collocations of selected frequent prepositional patterns in order to reveal their semantic prosodies and preferences (Partington, 2004). Left and right collocations are identified separately to pinpoint potential distinctive patterns bound either to the element modified by the prepositional phrase, or in turn to the prepositional complement.

First, prepositional patterns are identified and described within Czech newspapers, where the parameter of lemmatisation – newly introduced in this case study – is expected to be relevant due to the morphological variability of Czech. Next, an analogous study is conducted on English data. Subsequently, the results are summarised and viewed from two perspectives: to describe the prepositional phraseology of newspapers based on both the Czech and English corpus, and then to compare the findings cross-linguistically. Lastly, the methodology employed in this case study is briefly evaluated in order to inform the following case study.

4.5 Prepositional patterns in Czech newspapers

This chapter analyses the phraseology of newspaper discourse using the SYN2009PUB corpus of Czech journalistic texts.

Within this corpus I focus on *v*[31] ('in') and *na* ('on'), which are the most frequent prepositions in the Czech 3-5-grams retrieved from SYN2009PUB. *V* is also the most frequent Czech preposition overall in the representative SYN corpora (Český národní korpus, 2016).

4.5.1 Data

Being part of the PUB series of the SYN family of corpora, SYN2009PUB is designed with respect to quantity, to contain large amounts of data; it is not intended as representative (i.e. not balanced as regards the distribution of either periods of time or newspapers). However, the majority is formed of "quality" or "broadsheet" newspapers.[32]

The SYN2009PUB corpus contains 844,881,368 tokens (717,156,997 tokens excluding punctuation).

The texts contained in SYN2009PUB were published between 1995 – 2007. The source media are mostly "quality newspapers" or magazines, including both national and regional newspapers; the collection also comprises 1 tabloid daily (*Blesk*) and 1 opinion-oriented online newspaper (*Blisty*).[33]

4.5.2 Method

Lemmatised 3-5-grams were extracted from the SYN2009PUB; all lengths were retrieved together. The cut-off frequency was set at 10,000. The resulting dataset contained 1,342 n-gram types and 41,692,583 n-gram

[31] Lemmatised *v* includes the variant *ve*, which occurs mostly when followed by a word beginning either with *v-* or *f-* (*ve vesmíru*) or a consonant cluster (*ve středu*) (Cvrček & al., 2015, pp. 340–341).

[32] Cf. http://ucnk.korpus.cz/struktura.php

[33] Newspapers were evaluated with the help of the typology of the Czech media landscape at *Mapa médií* http://www.mapamedii.cz/mapa/typologie/index.php (NFNZ, n.d.)

tokens.[34]

Before describing the annotation of the data, I will comment on the method adopted and explain the motivations for the particular parameters of the search, mainly including punctuation, n-gram length and lemmatisation.

Due to a large degree of their variability, all figures in the data were automatically replaced with the # placeholder to allow for numerical n-grams to be lumped together.[35]

The n-grams were retrieved from lemmatised data, including punctuation – but not allowing for n-grams to occur across a sentence boundary. Punctuation was included since the pilot study suggested it may be a relevant part of textual phraseological patterning (cf. 4.3, where the frequent occurrence of quotation marks pointed towards the importance of direct speech in newspaper data).[36]

The n-gram length was set between 3-5 to subsume a broad variety of n-grams, while bearing in mind their potential informational value with regard to identifying frequent sequences and examining how they function in context. The previous chapter (parliamentary debates) suggested that different n-gram sizes reveal different types of phraseologies, both formally (fragments/phrases/whole turns) and functionally (e.g. bigrams corresponded to deictic expressions; 3-grams – to terminology and

[34] By n-gram types I mean the individual sequences (e.g. *in the first, in the world, in the past* are the three top frequent n-gram types within the *in* patterns); while n-gram tokens are individual n-gram occurrences (e.g. the n-gram type *in the first* is represented by 64,338 n-gram tokens, or hits in the corpus).

[35] I would like to thank dr. Michal Křen and dr. Pavel Vondřička from the Institute of the Czech National Corpus for their kind assistance with retrieving the data.

[36] Furthermore, Malá et al. (2021) has indicated that including punctuation in n-grams – commas above all – helps reveal more realistically how specific patterns are involved in text structuring. In the study cited, including punctuation allowed for the identification of patterns containing subordinators and occurring around syntactic boundaries, introducing dependent clauses – e.g. *chvíle, kdy jsme* ('moment when we') (ibid.).

aboutness phrases; 7-and longer – to performative formulae). In the light of those findings, combining multiple n-gram sizes within one search is intended to retrieve a broad range of patterns. I have selected those n-gram sizes which yielded the largest numbers of recurrent n-grams in the previous study, except bigrams, which were found to carry little lexical meaning, and hence are of limited value for the present functional-semantic analysis.

Admittedly, the downside of combining different n-gram lengths is that it results in a number of n-grams overlapping, e.g. the 4-gram *v sobota od #* includes the 3-gram *v sobota od*, both represented as separate hits. These overlaps call for a cautious interpretation of the frequencies, bearing in mind numerous duplicities. On the other hand, comparing overlapping n-grams may serve as a useful prompt towards revealing interesting details about their distributional tendencies. To illustrate this, let us examine patterns found in the present data, containing the string of lemmata *v tento* ("in this"), summarised in table 5.

n-gram (lemmatised)	translation	freq
v tento den	"on this day"	53,565
v tento případ	"in this case"	33,568
být v tento	"be in this"	32,996
se v tento	"se-reflexive in this"	22,487
v tento souvislost	"in relation to this"	14,039
total		156,655

Table 5. Overview of n-grams containing the string *v tento*

The verbal n-grams containing *v tento* were selected, i.e. *být v tento* and *se v tento*[37] , and their right-hand collocations were examined, to identify the noun forming the prepositional complement of *v* . The

[37] The reflexive pronoun *se* invariably constitutes a part of Czech reflexive verbs.

collocates (window = 0L, 1R, arbitrary frequency limit 1,000) were ordered by logDice so as not to favour rare or exclusive collocates (Brezina, 2018, p. 70). Tables 3 and 4 also include MI and T-score values for comparison – the ordering of the results by MI proves the same as logDice, while T-score produces a different ordering. The top frequent collocates found in this search are presented in tables 6 and 7.

lemma	translation	freq	MI	T-score	logDice
ohled	"regard"	1,667	9.638	40.778	9.302
směr	"direction"	2,902	9.062	53.769	9.111
chvíle	"moment"	2,741	8.359	52.195	8.508
případ	"case"	4,038	7.418	63.174	7.697
den	"day"	3,715	6.627	60.334	6.935

Table 6. BÝT V TENTO ___: top frequent patterns

lemma	translation	freq	MI	T-score	logDice
souvislost	"relation"	1,012	8.828	31.742	8.287
den	"day"	5,602	7.773	74.504	7.543
případ	"case"	1,477	6.520	38.013	6.270

Table 7. SE V TENTO ____: top frequent patterns

Comparing table 5 (the top frequent *v tento* patterns) with tables 6 and 7 (the top frequent verbal *v tento* patterns), it becomes clear that all the top frequent*v tento* patterns interact and overlap. The verbal pattern *být v tento* overlaps with all three top frequent nominal patterns (cf. table 6): *být v tento + v tento případ*, etc. Looking at the core nouns in the nominal patterns, *souvislost* only collocates with reflexive verbs *(se v této souvislosti*), while *případ* and *den* combine with the copular/existential *be* as well. This greater versatility of use may explain why *v tento den* and *v tento případ* are more frequent compared to *v tento souvislost*. However, it may simply be caused by the lemma *souvislost* being considerably less frequent in the corpus: it occurs 99 times per million tokens, compared to *případ* with 715 i.p.m. and *den* with 1,138 i.p.m. This seems to be reflected in the T-score values, too: the highest T-score is found with *day,* the most frequent of the three nouns. T-score is sensitive to corpus size (Brezina, 2018, p. 276) and tends to emphasise collocates which are very frequent overall (Cvrček & Richterová, 2019).

Elsewhere, by contrast, with the help of collocations, the overlap between n-grams was found to be even greater than suggested by the n-gram types retrieved. For example, *v sobota od* ("on Saturday at") is mostly followed by a figure (represented by *#*, expressing time), as shown in table 8; another roughly 1,700 occurrences of this 3-gram combine with a different prepositional complement. However, a closer inspection of the right-hand-side collocates reveals that the remainder of occurrences are followed by a numeral word, such as *devět* ("nine"). The only exception was *ráno* ("morning"), which is still a temporal expression. Hence, these n-grams point towards a semantic sequence (Hunston, 2008) which could be represented as: *v [day] od [hour or time of day]*.

n-gram	freq
v sobota od	17,059
v sobota od #	15,314

Table 8. Example of overlapping n-grams: *v sobotu od*

Finally, lemmatisation was employed to subsume all variant morphological forms – this is relevant especially in Czech given its rich morphological paradigms. During subsequent analysis in the KWIC view, the word form frequency breakdown offered additional insight into the distribution of individual variants. For example, in *V TENTO DEN*, it is the plural form that is vastly predominant in the newspaper data (cf. table 9). The plural form pattern *v těchto dnech* ("in these days") is typically used to report on events currently in progress, cf. ex. 12; or also, unexpectedly, as a vague time marker referring to recent events, presumably where focus is on the event and the precise timing is not considered essential (ex. 13).[38] This analysis in context points to the potential downside of lemmatisation: it may obscure the differences between the uses of the individual formal variants, as seen in table 9.

word	freq
v těchto dnech	49,944
v tento den	2,841
v tyto dny	719
v tomto dni	42

Table 9. Lemma V TENTO DEN: word-form frequency breakdown

12) "Je prostě právě to období [...], kdy houby přilákají do hor více lidí," uvedl i Karel Palička z Horské služby Beskydy. Houbaři mohou v těchto dnech celkem snadno najít atraktivní druhy [...]

    "'Right now it´s this time of year [...] when mushrooms attract more people to the mountains', said Karel Palička from the Beskydy Mountain Rescue. In these days mushroom pickers can fairly easily find attractive specimen [...]"[39]

13) Nový bronzový odlitek vysoký 3,5 metru odhalí 17. listopadu na Churchillově náměstí na Žižkově bývalá britská premiérka Margaret Thatcherová, která v těchto dnech potvrdila svou účast.

    "A new bronze statue, 3.5 m in size, will be unveiled on 17 November in Churchillovo náměstí in Žižkov by former UK PM Margaret Thatcher, who has confirmed her attendance in these days."

[38] Another potential issue as suggested in the previous chapter is that of variable word order within patterns. This problem will be addressed in the following case study.
[39] All English translations of the Czech corpus examples are mine, unless specifically stated otherwise.

4.5.3 Prepositional patterns: annotation

Within the 3-5-grams, n-grams containing prepositions were isolated manually, yielding a total of 437 n-gram types and 11,512,598 n-gram tokens. In terms of tokens, prepositional n-grams form 27.6% of the whole 3-5-gram set. Each prepositional n-gram was annotated with the preposition it contained.[40] A total of 19 prepositions were represented. This annotation by preposition was then used to sort the n-grams. The resulting groups are summarised in table 10, ordered by n-gram type frequency.

As seen in table 10, the representation of prepositions in the n-grams roughly corresponds to the frequency list of individual prepositions: most of the 20 top frequent prepositions in the corpus overall are also found in the top frequent prepositional 3-5-grams.[41] Compared to the reference frequency lists based on the representative SYN corpus family (Český národní korpus, 2016), the results correspond with regard to *v*. In all these reference corpora *v* proves to be the most frequent Czech preposition, and ranks among the top frequent 4 lemmata (Český národní korpus, 2016).

Being the most frequent prepositions in n-gram types, *v* and *na* will be analysed in detail with regard to their phraseological environments. Further, *v* allows for a subsequent comparison with its English translation equivalent *in*, which is discussed in the English newspaper study in subchapter 4.6.

[40] Where an n-gram contained two prepositions (no n-grams contained more than two), it was annotated for both but sorted according to the first preposition: e.g. *od # hodina v* ("since # o´clock in") was annotated as *od, v* and included in the *od* category. Typically, multiple-preposition n-grams were sporadic and constituted a minority of the category (represented by one or two n-gram types only), with the following three notable exceptions:

- *v – od* (*v sobota od #,* 'on Saturday from #') 5 n-gram types
- *v - s* (*v čelo s*, 'led by') 6 n-gram types
- *od – do* (*od # do #,* 'from # to #') 6 n-gram types

[41] While the few remaining prepositions (marked in bold) did not occur in the n-grams, they may well occur in less frequent n-grams, below the arbitrary cut-off frequency employed here.

prepositional 3-5-grams			prepositions alone		
rank	preposition	n-gram tokens	rank	preposition	tokens
1	v	4,709,905	1	v	21,429,414
2	do	1,579,530	2	na	13,908,323
3	na	1,438,233	3	z	7,139,781
4	o	957,886	4	s	6,230,222
5	od (od-do)	821,212	5	do	5,207,684
6	z	473,773	6	o	4,773,606
7	s	389,250	7	k	3,096,617
8	za	277,653	8	za	2,980,512
9	pro	221,535	9	pro	2,776,913
10	k	173,234	10	po	2,611,257
11	před	116,760	11	od	2,173,661
12	podle	95,816	12	podle	1,469,198
13	po	81,172	13	u	1,360,155
14	při	65,548	14	před	1,270,891
15	nad	42,393	15	při	1,143,932
16	vzhledem k	32,513	16	mezi	774,020
17	u	13,463	17	nad	717,007
18	kvůli	11,701	18	bez	575,674
19	mimo	11,021	19	proti	559,252
			… 22	kvůli	357,843
			27	mimo	179,042
			31	vzhledem k	97,680

Table 10. SYN2009PUB prepositional n-grams by n-gram token frequency, compared with the overall frequency list of prepositions in SYN2009PUB

4.5.4 *V* prepositional patterns: analysis

Patterns containing *v* were the most frequent prepositional pattern group (4,709,905 n-gram tokens, falling under 170 n-gram types).

First, *v* n-grams were annotated based on whether they contained a left-side pattern, i.e. preceding and postmodified by the preposition *v* (= L); or a right-side pattern following *v,* i.e. its prepositional complement (= R); or whether they spanned both sides of the preposition (LR). The aim was to identify which type of patterning around the preposition was more frequent. Also this annotation serves to indicate whether examining left or right collocations of the whole patterns would be more informative.

The results of this annotation are summarised in table 11. The right-hand side patterns form the majority (roughly 66%), indicating that *v* forms the most recurrent chunks with its prepositional complements, although left-hand side patterns are not marginal, forming about 25% of the total. I will briefly address each of these three pattern groups with regard to its composition. An interesting finding was revealed by comparing the numbers of n-gram types and n-gram tokens: their ratios are very similar for both the left- and right-hand side patterns, suggesting a similar degree of lexical richness on either side of the *v* patterns.

side	L	LR	R	total
example	mistrovství svět v “world championship in”	dnes v # hodina “today at # o´clock”	v první polovina “in the first half”	
n-gram types	37	24	109	170
n-gram tokens	1,179,492	414,406	3,116,007	4,709,905
n-gram type-token ratio	3.14	5.79	3.5	

Table 11. Frequencies of left- and right-hand side patterning around *v*

Another layer of annotation (applied across the left/right distinction) was based on semantic/functional criteria. Patterns were characterised bottom-up according to formal characteristics (e.g. parts of speech contained in the pattern), the lexical meanings they conveyed or textual functions they fulfilled. This categorisation was corpus-driven (Tognini-Bonelli, 2001) by the particular dataset, i.e. no categories were preconceived.

Among the left-hand side patterns (table 12 below), the largest group were relative patterns such as *, který být v* "which be in", introducing a relative clause. The second most frequent group were patterns containing conjunctions: coordinating (*, ale i v "*, but also in") or subordinating (*ten, že v* "the fact that in"). In sum, most left patterning around *v* is found at syntactic boundaries. The next frequent groups of left patterns (in descending order of n-gram token frequency) were nominal and verbal. Both these groups were repetitive, limited to few n-gram types (5 nominal, 6 verbal). The nominal patterns referred to institutions and places (*krajský soud v* "regional court in"), one was temporal (*# hodina v* "# o'clock in"). The verbs were verbs of happening (*se uskutečnit/konat/hrát/stát v* "take place/be played/happen in"), and one verb of saying, introducing direct speech (*, " říci v* "say in").

formal tag	n-gram token freq	example
relative pronoun	455,512	, který být v
conjunction	284,230	ten , že v / , ale i v
noun	128,040	mistrovství svět v
verb	120,443	se uskutečnit v
be	92,974	ten být v
numeral	51,029	. # . v
punctuation	29,140	) , v
pronoun	18,124	, a ten v
total	1,179,492	

Table 12. L *v* patterns

The patterns spanning both sides (LR, table 13) all convey temporal meanings. Most frequently they contain a noun as the prepositional complement of *v* (*dnes v #* hodina "today at # o'clock"); the majority also contained the verb *be* (*být v současný doba* "be currently"). Others contained verbs of happening, often combined with a numeral referring to a point in time (*začínat v # hodina* "start at # o'clock").

formal tag	n-gram token freq	example
be + NP	156,195	být v současný doba
noun	100,894	dnes v # hodina
verb of happening + numeral	64,205	začínat v # hodina
conjunction (all coordinators)	60,080	a v rok #
numerals (subsume some NP´s)	33,032	v # hodina v
total	414,406	

Table 13. LR *v* patterns

The right side patterns (table 14) were the most numerous (66% of the total), but not diverse in terms of form or function. The majority were again combinations of *v* with a noun phrase complement, conveying adverbial meanings: temporal (*v tento den* "on this day"), place (*v kulturním domě* "in the culture hall") or circumstances (*v pořádku* "in order"). Further patterns included numerals (some also temporal in meaning), relative clause openings and sequences with pronouns, whose textual functions vary.

formal tag	n-gram token freq	example
noun	2,396,425	v tento den
numeral	385,638	. v #
relative pronoun	175,903	, v který být
pronoun	158,041	v on být
total	3,116,007	

Table 14. R *v* patterns

Since the right-hand side patterning is prevalent, in the analytical section of this subchapter I will examine *v* patterns focussing predominantly on their right collocations. The next section describes the annotation procedure in more detail.

4.5.4.1 Notes on annotation

The *v* n-grams were further manually annotated and sorted based on two types of criteria: formal and functional. The formal annotation was essentially based on the parts of speech contained in each n-gram type, with some tags being slightly more narrowly conceived than that; this will be explained below. These tags subsumed a specific subtype of some parts of speech which were found to occur frequently in the results and therefore seemed to warrant a closer analysis.

The reflexive pronoun clitics *si* and *se* were disregarded during annotation since they are not independent words – they form parts of reflexive verbs and the verbal element was never present in the n-grams, therefore the reflexive pronoun was viewed as a fragment. All other function words were annotated: these comprised conjunctions and relative pronouns. N-grams where *v* was the only word, accompanied by punctuation (e.g. *) , v*), were annotated with a special tag (*punct*), since punctuation is expected to be involved in patterns contributing to text structure. N-grams containing verbs were annotated using two tags: *be* for those with the verb *be* and *verb* for all other verbs (most were lexical verbs).

The tag *noun* subsumes noun phrases too, i.e. *městský úřad v* was marked up as *noun* (with no separate tag for the adjectival noun modifier). On the other hand, where a noun was preceded by a numeral, these numerals were conceived as a separate annotation category: this was because numerical expressions were very frequent in this particular prepositional pattern group. For the same reason, relative/interrogative pronouns were conceived as a separate category because they occurred frequently. This was another case where punctuation proved to be involved in patterns, since phrases formed by *v +* relative/interrogative pronoun are obligatorily preceded by a comma, e.g. *, v který.*

Where applicable, the formal annotation tags were combined so that each n-gram was matched with a part-of-speech-gram of sorts, for instance: *začínat v # hodina* "start at # o'clock" = *verb num noun*.

Table 15 lists the frequencies of individual formal groups ordered by n-gram token frequencies, though this roughly corresponds to the ordering by the type frequencies. This table presents only a rough overview; in case of multiple tags, only the first tag is counted, e.g. *začínat v # hodina* "start at # o'clock" = verb num noun – counted as *lexical verb*.

formal tag	n-gram types	n-gram tokens
noun	94	2,091,117
numeral	20	1,003,941
relative	14	631,415
conjunction	11	344,310
be	13	249,169
lexical verb	9	184,648
pronoun	7	176,165
punctuation	2	291,40
total	170	4,709,905

Table 15. Formal classification of *v* patterns, ordered by n-gram token frequency

The next step was a semantic/functional annotation. Each *v* pattern was characterised with regard to its form and/or semantics (where identifiable, e.g. where the pattern contained a lexical verb such as a verb of saying: *, " říci v*), or textual function (e.g. patterns containing a coordinating conjunction: *# a v*). Again, tags were combined to provide a detailed description of each pattern for the purposes of subsequent qualitative analysis, e.g. *začínat v # hodina* "start at #o'clock" = *happening,* *temp* (verb of happening + temporal adverbial element). The resulting classification is summarised in table 16 below, ordered by the frequency of n-gram tokens.

tag	example + translation	n-gram types	n-gram tokens
temporal	dnes v # hodina “today at # o'clock”	65	1,840,384
relative/interrogative	, který být v “, which be in”	14	631,415
numeral	v výše # milión “worth # million”	14	537,790
place	v centrum město “in the city centre”	23	421,712
circumstances	v domácím prostředí “in domestic environment”	14	266,387
content clause	, že být v “, that be in”	3	243,571
verb of happening	se uskutečnit v “take place in”	8	173,575
be	být v tento “be in this”	8	155,876
secondary preposition	v souvislost s “in relation to”	5	130,202
pronoun	v tento “in this”	4	126,067
coordinator	a v rok # “and in year #”	8	115,884
classification	v jeden z “in one of”	1	23,843
subordinator	, aby v “, so as to in”	1	21,461
verb of saying	, " říci v “, " say in”	1	11,073
other (unclassified)	) , v “) , in”	1	10665
total		170	4,709,905

Table 16. Semantic/functional classification of *v* patterns, ordered by n-gram token frequency

Note that in case of multi-part tags such as *happening, temp* etc., these were counted according to the first tagged element, so that there is a certain degree of overlap between the semantic pattern types. The frequencies are therefore only for general reference – the tagging allowed for a more fine-grained analysis taking those overlaps into account.

A comparison of the n-gram type and token frequencies in table 16 suggests that some semantic groups of *v* patterns comprise a relatively more varied repertory than others. This seems to be linked with whether a given pattern group conveys lexical or grammatical meanings. Temporal or spatial patterns, for instance, seem to be a fairly varied group, referring to a variety of settings. On the other hand, patterns introducing dependent clauses (relative/interrogative, content clause), albeit frequent, are more repetitive in terms of types, probably because they are more formally conventionalised (there is the obligatory comma and a limited repertoire of pronouns which may be contained in these patterns) – this corresponds to the fact that they serve a specialised grammatical function.

4.5.5 Analysis of selected *v* pattern groups

Below, two selected semantic/functional groups of *v* patterns will be introduced in more detail, with regard to their semantic preferences and prosodies. One group is defined semantically (temporal patterns), while the other functionally (secondary prepositions). Temporal patterns were selected based on their frequency (cf. table 15), while secondary prepositions are an opportunity to extend the focus of this study on prepositional patterns.

4.5.5.1 Temporal patterns

The temporal group was by far the most frequent *v* pattern category (65 n-gram types and 1,840,384 n-gram tokens). As regards the part-of-speech composition of these patterns, the vast majority comprise nouns or noun phrases (*v tento den* "on (lit. in) this day"), or combinations of nouns and numerals (*v # hodina* “at # o’clock“).

In the following part I examine the collocations of the top frequent temporal patterns comprising v in any position, to explore their potential semantic prosodies and semantic preferences. As my formal annotation has indicated, there is clearly more patterning on the right-hand side of the *v* patterns overall; the same is true of the temporal *v* patterns (formally comprising 2 n-gram types L, 11 LR, 52 R), hence I focus on right collocations. Table 17 lists the top 10 most frequent right *v* temporal patterns.

pattern	translation	freq (n-gram tokens)
v # hodina	“at # o’clock”	204,595
v # . minuta	“in #. minute”	158,569
v současný doba	“in current time”	101,009
v rok # .	“in the year #.”	83,157
v # hodina .	“at # o’clock .”	71,767
v tento den	“on this day”	53,565
v příští rok	“in the next year(s)”	43,923
v sobota #	“on Saturday #”	40,094
v loňský rok	“in the last year”	40,001
v doba ,	“at the time”	37,646
total top 10		834,326

Table 17. Top 10 most frequent right *v* patterns

There is considerable overlap between the patterns in terms of the words included – the same holds for the whole list, not only the top ten. Recurring lemmata include temporal nouns (*hodina, minuta, den, rok, doba*) and names of days (*sobota, neděle*). Apart from these, there are discourse-specific expressions related to time periods in sports: *první kolo* (“first round”)*, první poločas, první půle* (“before half-time, first half“).

Based on the top frequent n-grams, I have selected the following sample of temporal patterns around which the collocation queries are centred:

- v ADJ rok | v rok NUM, including *v příští/loňský rok* and *v rok #*
- v (___) doba (___) to include both *v současný doba* and *v doba(, kdy)*
- v ___ hodina
- v ___ minuta

The collocation search parameters were set as follows: collocation window between R1 and R3;[42] minimum frequency in the corpus = 100, minimum frequency of the collocate in this particular context = 10; sorted by logDice.

- v ADJ rok ("in ADJ year") | v rok NUM ("in year NUM")

The top collocate of this pattern was the relative pronoun *kdy* "when", introducing a relative clause specifying which year is being referred to. Typically, the year is further characterised using a verb of happening, referring to an event which took place in a particular year. Verbal collocates conveying this included *získat* ("obtain"), *stát* ("happen"), *založit* ("establish"), *být* ("be"), *začít* ("begin"), *dosáhnout* ("achieve"), *vydat* ("publish").

14) Režisér za něj v loňském roce získal prestižní cenu Alfreda Radoka
"Last year the director was awarded the prestigious Alfred Radok prize for it"

- v (___) doba ("in ___ time")

The most frequent n-gram with *doba* was *v současná doba* – this variant constitutes 39% of all the *v __ doba* hits in the corpus. The query

[42] The collocation span is set as following and excluding the last position of the n-gram.

also allowed for other premodifiers of *doba*, which were for the most part similar adjectives: *současná* ("current"), *poslední* ("recent"), *blízká* ("near"), *tato* ("this"), *dnešní* ("today´s"), *dohledná* ("foreseeable") etc. Overall, the focus tends to be on the present time, with some extension to either the near future or recent past. This corresponds to the orientation of newspaper articles towards current events.

The top three collocates are the verbs *probíhat* ("be in progress"), *připravovat* ("prepare"), *pracovat* ("work"); further frequent verbs were *žít* ("live")*, dařit* ("succeed, thrive")*, jednat* ("negotiate"), *existovat* ("exist"), *působit* ("be active"). These collocates suggest that *v ___ doba* refers to activities currently in progress, under preparation, or states potentially subject to change (e.g. *currently living/working* somewhere, etc.).

15) Ani Maďarsko nechce zůstat stranou, a proto v současné době připravuje rozsáhlý program k povzbuzení investic.
"Not wanting to lag behind, Hungary is also currently preparing an extensive investment incentive programme."

Other than verbs, there were the temporal or degree adverbials *již* ("already")*, stále* ("still"), *už* ("already")*, často* ("often"), *hodně* ("a lot"), indicating that *v ADJ doba* collocates with focussing or intensifying expressions.

- v ___ hodina ("at ___ o'clock") X v ___ minuta ("in ___ minute")

Though these patterns seem similar at first sight, their collocational characteristics are markedly different.

V ___ hodina (in fact the word form is typically in the locative case, ___ *hodin* "at ___ o'clock") is followed either by a specification of the time of day (*ráno, odpoledne, večer, dopoledne* "morning, afternoon, evening, morning"), a venue (*sál, kostel, náměstí, kino, divadlo* "hall, church, square, cinema, theatre"), or type of event (*koncert* "concert", *výstava* "exhibition"); also the related noun *vstupné* ("admission fee"). Verbal collocates pertain to the same semantic field: *zahájit* "open", *začínat "begin", vystoupit "perform".* To sum up, *v ___ hodina* shows a semantic preference for specifying the circumstances of cultural or other events.

Where the word form is other than locative, these patterns usually refer to times of the day, e.g. *v ranních/nočních hodinách*. ("in morning/night hours"). The use of the plural form here seems typical of journalism, as testified by results from the SYN8 corpus (table 18).

text-type	freq	i.p.m.
journalism	103,175	21.29
non-fiction	2,812	6.85
fiction	621	4.64

Table 18. Distribution of *v ___ hodinách* across text-types in SYN8

On the other hand, *v ___ minuta* (typically it also takes the locative case, *v ___ minutě* or plural *v ___ minutách*) is prominent in sports reports. Collocates include *nastavení, prodloužení* ("overtime"), *prohrávat* ("lose"), *vyrovnat* ("even up"), *inkasovat* ("receive"), *skóre* ("score"), *půle* ("half"), *třetina* ("third"), *poločas* ("half-time"). All the major collocates clearly suggests references to sports matches, with the exception of *dějství* ("act"); however, a look at the concordances reveals that these instances of *dějství* are likewise references to parts of a sports match. Some of these instances are accompanied by other highly marked stylistic choices, representing a sports match in an exalted poetic style, as in ex. 16, where the ice hockey goal is conceptualised as a "shrine"; other marked expressions are highlighted in bold.

16) V poslední minutě úvodního dějství nastřelil Skuček tyč Bláhovy svatyně. V první minutě druhé části orazítkoval domácí Komárek pro změnu břevno branky hostí.

"In the final minute of the opening act, Skuček shot the post of Bláha´s shrine. In the first minute of the second part, the local

Komárek stamped the guest team´s goalpost in return"

4.5.5.2 Temporal pattern collocations – interim summary

Temporal patterns were found to be lexically quite repetitive, as a large number of lexemes were shared across patterns – mostly temporal nouns (*hodina, minuta, den, rok, doba*, *sobota* "hour/o'clock, minute, day, year, time, Saturday"). Other recurrent lexical elements were discourse-specific, occurring in sports sections, e.g. *první poločas* ("first half").

The collocation analysis (span between 1R and 3R) was focussed on patterns containing the generic temporal nouns (*rok, doba, hodina* and *minuta*), in order to select patterns likely to occur across different newspaper sections, and to avoid a potential bias towards particular topics, newspaper sections or discourse types (e.g. sports terminology). This strategy worked with patterns containing *rok* and *doba*. *Rok* patterns usually refer to particular points in the past, while *doba* more commonly refers to events in progress over the present time, reporting on current affairs. *Doba* also co-occurs with focussing expressions *již, hodně*, which attract the reader´s attention and may be used by the journalist to present their interpretation, to assign importance to particular events etc.

However, even some of these seemingly neutral patterns show a semantic preference. *V ___ hodina* usually conveys information about a planned event, while *v ___ minuta* occurs predominantly in sports reports. Some of the collocates of the pattern *v ___ minuta* were stylistically marked, suggesting a tendency towards embellishing the language of sports reports and making it compelling to the reader.

4.5.5.3 Secondary prepositions

Since the methodological tenet of this case study is using function words as the basis for examining phraseological patterns, it seems relevant to pay attention to those phraseological sequences involving the preposition *v* which themselves fall under "function patterns" rather than "content patterns". Hence, in this section I examine the collocations of *v* patterns which form secondary multi-word prepositions.

These prepositional *v* patterns were represented by five n-gram types. Table 19 lists the prepositions in their word form rather than as lemmatised n-grams, since their form is fixed and therefore is the only manifestation of the lemmatised n-gram.

preposition	translation[43]	freq (n-gram tokens)
v souvislosti s	" regarding/concerning"	49,197
ve srovnání s	"compared to"	26,353
v čele s	"led by"	20,770
v souladu s	"in accordance with"	20,725
v rozporu s	"contrary to"	13,157
total		130,202

Table 19. Secondary prepositions with *v* by frequency

- *v souvislosti s* ("in connection with; regarding/concerning")

The top-ranking collocates *chystaný*, *připravovaný* "under preparation", *blížící* "approaching", *výstavba* "construction" hint towards an association with plans and future outlooks. Other collocate nouns *privatizace* "privatisation", *reforma* "reform", *ukončení* "termination", *novela* "novelisation", *rekonstrukce* "reconstruction/refurbishment" correspond to this, referring to actions which bring about a major change of affairs.

Apart from this, remarkably, most collocates of *v souvislosti s* clearly point towards a negative semantic prosody: cf. the nouns *aféra*

[43] Translation equivalents were found in the InterCorp parallel corpus via the Treq application (treq.korpus.cz), with the exception of *in accordance with,* where I used my own judgment since the Treq results did not seem plausible.
Cf. also the translation of *v souladu s* as *in accord(ance) with* in https://slovniky.lingea.cz/anglicko-cesky/v%20souladu%20s.

"affair/scandal"*,* *kauza* "case"*,* *krach* "bankruptcy/crunch, breakdown"*,* *atentát* "assassination"*,* *útok* "attack"*,* *vražda* "murder"*,* *krize* "crisis"*,* *povodeň* "flood"*,* *skandál* "scandal" and adjectives: *korupční* "corruption, bribery"*,* *teroristický* "terrorist"*.* The noun *vyšetřování* "investigation" also seems associated with (at least potentially) negative phenomena.

While the noun *souvislost* "connection" is in itself emotionally inexpressive, the secondary preposition *v souvislosti s* shows a distinct semantic preference for negative collocates. These findings confirm the notion that a word´s negative semantic prosody is often not identifiable solely on the basis of the word itself – it is only revealed through an analysis of the word´s recurrent contexts (cf. Sinclair´s classic examples of *happen* and *set in,* Sinclair 1987, in Partington, 2004, p. 132; or *budge,* Sinclair, 2004, pp. 142–147)*.*

- *ve srovnání s* ("compared to")

Ve srovnání s obviously introduces comparisons between entities or developments. It collocates mostly with adjectives referring to the past: *loňský* "last year´s", *předešlý, předcházející, minulý* "previous, preceding", as well as nouns referring to points in time: *čtvrtletí* "quarter", *pololetí* "half-year", and numerals referring to specific years such as *1998.* Related to these are words describing development over time - *poklesnout* "decrease", *stoupnout* "increase"*,* *vzrůst* "growth", *průměr* "average". Yet another group of collocates describes entities against which the comparison is being made: *ostatní* "other", *vyspělý* "advanced", *okolní* "surrounding", *konkurence* "competition", *jiný* "other", *týž* "the same".

An interesting collocate was *postkomunistický* "post-communist", suggesting comparisons of different geographical regions – presumably a frequent subject of discussion in the news during the 1990s, given these countries´(including Czechia´s) recent history. A related collocate is *vyspělý* "advanced", serving to compare (and simultaneously evaluate) countries. Indeed, lower-ranking collocates include *Maďarsko* "Hungary" and *země* “country”. These collocates are yet another example of how recurrent patterns reflect frequent topics present in the newspaper corpus texts, and by extension allow a glimpse into the topics which had a salient presence in the news discourse of a particular period.

- *v čele s* (“led by”)

The right-hand collocates of *v čele s* essentially serve the function of introducing an agent, typically a concrete person. Collocates included nouns describing professions or other positions (*frontman*) and proper names (*Jágr*). These collocates prove informative as to what persons are represented as holding a position of some power – although as will become clear from the individual examples, the nature or real impact of said power may vary greatly. They can be grouped into the following major clusters:

- artists (*frontman, režisér* “director”, *dirigent* “conductor”),
- sportsmen (*Jágr*, *brankář, gólman* “goalkeeper, goalie”),
- politicians (*hejtman* “governor”, *primátor* “mayor”, *premiér* “PM”, *Arafat*),
- (marginally) members of the army (*kapitán* “captain”, *generál* “general”).

Interestingly, sports-related patterns occur yet again, again suggesting the prominence of sports reporting in newspapers.

These agent nouns pose an opportunity to examine differences in the frequency of gender representation in leading positions, since Czech nouns are marked for gender. However, if feminine variants of nouns are present at all (within the top 50 collocates only three instances were found, see below), they are considerably less frequent that their masculine counterparts:

- *brankář 509, brankářka 37 -* “goalkeeper”; (+ note additional instances of the masculine synonym *gólman* 220)

- *předseda* 561*, předsedkyně* 67 “chairman, chairwoman”
- mistr 207, mistryně 37 “champion”

Using the *Corpus Calculator* “2 words in 1 corpus” tool (Cvrček, 2021), with regard to corpus size, the frequency differences in all these three cases were acknowledged as significant by both Chi2 and LogLikelihood statistical tests (level of significance was set at 0.05.) This suggests that women leaders are less frequently depicted than male ones in Czech media, which corresponds to the underrepresentation of female speakers identified in British news from roughly the same period, as described in section 4.3 of the earlier pilot study.

As regards semantic preference, most collocates were descriptive agent and proper nouns, yet a few other collocates point towards a tendency of *v čele s* towards a positively evaluative prosody: *charismatický* “charismatic”, *legendární* “legendary”, *vynikající* “outstanding”, *mistr, mistryně* “champion”. A complementary look at the left-hand collocates of *v čele s* confirmed this tendency: *špička* “top”, *zručný* “skilled”, *elita* “elite”, *hvězda* “star”. (Other than these positively evaluative expressions, the left collocates corresponded to the right-hand ones: they comprised words referring to groups of people *delegace* “delegation”, *porota* “jury”, *průvod* “parade”, *konsorcium* “consortium”, whose leader is then duly introduced by *v čele s*.)

- *v souladu s* (“in accordance with”)

This preposition shows a clear preference for legal contexts. Collocates include *předpisy* “regulations”, *ústava* “constitution”, *legislativa* “legislation”, *norma* “norm” and other nouns and adjectives from the same semantic field. Outside the legal context, it collocates with *trendy* “trends” ,*očekávání* “expectations”, *požadavek* “requirement” or *plán* “plan”. Overall, *v souladu s* seems to have an implicitly positive semantic prosody, in keeping with the lexical meaning of *soulad.*

- *v rozporu s* ("contrary to")

V rozporu s represents the opposite meaning than that of *v souladu s* and this is reflected in the two prepositions sharing a number of similar collocates from the legal field: *předpisy* "regulations"*, ústava* "constitution", *legislativa* "legislation". Further collocates were *etika* "ethics", *logika* "logic", *princip* "principle", *tvrzení* "assertion", *rozkaz* "order" or *mravy* "morals", which in fact ranked first*,* representing the fixed phrase *v rozporu s (dobrými) mravy* "contrary to good manners". We can conclude that *v rozporu s* manifests a negative semantic prosody, implied by the lexical meaning of *rozpor.*

In summary, the collocates of secondary prepositions containing *v* have revealed that each complex preposition has its distinct semantic preference and/or prosody. In some cases, the evaluative prosody seems directly derived from the lexical meaning of the noun contained in the prepositional sequence, as in *v souladu/rozporu s.* Elsewhere, the semantic prosody was opaque, as with *v souvislosti s,* found to systematically occur in negative contexts. Further, as regards semantic preference, some secondary prepositons tend to occur in particular types of register (*v souladu/rozporu s* collocated with legal settings) or thematically defined contexts (*v čele s* is used to introduce personal agents). A diachronic perspective on the collocations of patterns would be an interesting next step, to reveal whether and how the semantic prosodies of patterns may evolve over time.

4.5.5.4 Summary: *v* prepositional patterns

V was the top frequent preposition and *v* patterns were the most frequent prepositional pattern group altogether in the newspaper corpus.

The patterns were annotated formally (by parts of speech involved in the patterns) and semantically/functionally. The formal annotation revealed that the most frequent POS tokens were nouns and numerals, constituting prepositional complements; followed by relative pronouns and conjunctions, suggesting that patterning tends to recur around syntactic boundaries and may be linked to conventionalised ways of linking clauses.

The semantic annotation was applied where the pattern contained a lexical word; it was based functionally where the pattern did not clearly convey a particular lexical meaning but instead served a textual function, such as patterns containing a conjunction. Judging by type and token frequencies, the patterns conveying lexical meanings were found to be relatively more formally diverse, while the functional patterns were more formally conventionalised, adjusted to a particular grammatical of textual function.

Most often, recurrent patterning was found on the right-hand side of the preposition *v* – for this reason, the following collocation analysis was centred around right-hand collocates. Confirming the suggestions of the POS annotation, the right patterning around *v* typically consists of a noun phrase forming the prepositional complement and conveying adverbial meanings (temporal, spatial, or of other circumstances). Most left patterning around *v* was linked to syntactic functions – introducing coordinated or subordinate clauses, predominantly relative ones.

In a qualitative collocation analysis, two *v* pattern groups were examined: one lexical (temporal adverbial patterns) and one functional (secondary prepositions). Lexical patterns are characterised by a conspicuous presence of recurrent general temporal nouns such as *hodina, minuta, den, rok, doba*. A collocation analysis revealed that while some of these general temporal nouns are involved in patterns conveying functions which are predictable in the newspaper register (referring to past events or to events currently in progress), others are typical of particular newspaper discourse types (e.g. *v___ minuta* restricted to sports reports); in this case the collocation analysis also helped reveal some unexpected style characteristics of the sports report discourse.

The *v* secondary prepositions manifest distinct semantic prosodies, which are sometimes transparent (linked to the lexical meaning of a noun co-forming the preposition), but sometimes opaque (*v souvislosti s* recurs in negative contexts). Some prepositions also show a preference for particular registers or contexts (e.g. *v souladu s* in legal contexts).

4.5.6 *Na* prepositional patterns: analysis

In this section, devoted to the second most frequent preposition *na*, I have opted for a different approach than in the analysis of *v* patterns. This was motivated by the quantity of the material: there were much fewer n-gram types containing *na* (60) than *v* (170), making the *na* dataset easier to work with. Hence, looking at the *na* n-grams, several functional pattern types were readily observable: there was no need for an annotation as sophisticated as that employed in the *v* analysis. In this section, I will briefly characterise the patterns found with *na*, and proceed towards a functional classification and description of the salient functional types of *na* patterns.

The top 10 most frequent *na* patterns comprise 6 containing a comma, pointing towards recurrent patterning around clause boundaries – specifically relative or content subordinate clauses, echoing the same tendency which was observed in *v* patterns. At the same time, the top frequent *na* patterns do not lend themselves to a semantic or functional classification: they mostly contain function words, cf. the top 5 list below:

1. na ten , ("on that")
2. , že na ("that on")
3. , který na ("which on")
4. , na jenž ("on which")
5. *na ten , že* ("on the fact that")

Apparently, these sequences are very frequent precisely because they are multifunctional and do not convey any very specialised meanings.

Furthermore their frequency may be enhanced by the formally stable components of the pattern: the punctuation and the finction words (pronouns or the conjunction *že*).

Two miscellaneous observations were made examining the *na* pattern frequency list. First, remarkably, *na jenž* 'on which' (n-gram token freq = 71,766) ranks higher than *na který* (specifically *na který být*, 23,138 hits). Both these patterns are synonymous and introduce a relative subordinate clause, differing only in style: the relative *jenž* is found predominantly in written language and considered stylistically marked, more formal than its neutral counterpart *který* (Cvrček & al., 2015, pp. 27; 267; 367). This suggests that the higher frequency of *jenž* may be determined by the written medium of this corpus. However, this does not seems to be the case: the lemma *jenž* alone occurs 1,442 i.p.m. in the SYN2009PUB corpus, while *který* is much more frequent a 7,103 i.p.m. Furthermore, the lemma sequence *na jenž* occurs 73,792 times and 83 times per million tokens (henceforth i.p.m.) in the corpus, while *na který* is again more frequent: 98,115 times and 116 i.p.m. A chi-square test confirms that this difference is statistically significant (P = 0.01932) (Cvrček, 2021).[44] Hence *jenž* may be preferred over *který* only in particular patterns, perhaps the attributive patterns *na jehož, na jejichž* etc. ("on whose", revealed by the frequency breakdown of *na JENŽ*), which have no counterparts containing *který*.

Secondly, the pattern *nový město na* is conspicuous, corresponding to the place name Nové Město na Moravě, a small town not associated with any major economic or political significance. Its presence among the top frequent n-grams is probably partly caused by its length – other cities are undoubtedly mentioned more frequently but their contexts are likely to vary, hence they would occur in a number of different n-grams. Also it may

[44] Test statistic = 5.4724, level of significance = 0.05, "2 words in 1 corpus".

be due to the regional dailies *Deníky Moravia* being included in the corpus. Finally, this may be influenced by the fact that this town hosts major sports events.

In the following part I will describe the most frequent groups of patterns identified in the *na* pattern group, listed in table 20.

pattern group	example + translation	n-gram types	n-gram tokens
numerical data	na # tisíc "around/at # thousand"	12	296,022
aboutness words	na mistrovství svět; (na snímek) "at the world championship"; "(in the photo)"	13	199,443
verb + O_{PREP}	záležet na tom "depend on it/depend on the fact whether"	7	97,988
text structuring	na první pohled "at first sight"	3	44,802
journalistic style	moci těšit na "can look forward to"	3	35,313
other		22	764,665
total		60	1,438,233

Table 20. Functional groups of *na* patterns

With numerals, *na* conveys two meanings: either the selection of the preposition is determined by the valency of a preceding verb (*odhadovat na* "estimate at"), or the preposition expresses approximation[45] (SSJČ, 2011) (ex. 17).

17) v Praze natrvalo žije na sedm tisíc Američanů

[45] SSJČ: *na* „2. u číslovek vyjadřuje přiblížení se k prostorové n. časové hranici, přibližnost číselného n. časového údaje: [...] *sněhu je na dva metry; kapesníků má na tucty"*
("*na* with numerals expresses approximation to a spatial or temporal limit, approximate numerical or temporal information")
available from
https://ssjc.ujc.cas.cz/search.php?heslo=na&sti=39846&where=hesla&hsubstr=no

"there are around 7000 Americans permanently residing in Prague"

Patterns containing aboutness words were varied: some referred e.g. to sports (*na druhém místě* "in the second place", *na mistrovství světa* "at the world championship"). Other content patterns carried spatial meaning, specifically serving extratextual reference, referring the reader to external sources of information: *na telefonní číslo* "at the telephone number", *na snímku* "(pictured)", *na internetových stránkách* "on the website".

The remaining pattern groups were represented by relatively few examples. Worth a comment is the group of patterns which were found to be typical of journalistic style: *na svém kontě* "to his/her name", used metaphorically to introduce a person´s achievements or past actions. The evaluative meaning conveyed by this pattern can be either positive or negative depending on the context, as attested by 18, 19; some uses are expressive and ironic, cf. 20.

18) Na svém kontě má řadu hitů
"He has a number of hit songs to his name"

19) Dva svícny z černého kovu odnesl neznámý pachatel z hrobky na kladenských hřbitovech. Na svém kontě tak má škodu za 40 tisíc korun.
"An unknown intruder has taken two candleholders from a tomb at the Kladno graveyard. As a result he has a damage worth 40 thousand Czech crowns to his name."

20) Neplnění pracovních povinností, porušení nařízení rady! Takové prohřešky má na svém kontě místostarosta Prahy 13 [...]
"Failure to perform his duties and to comply with the Board´s orders! That is the sort of misconduct the vice-mayor of Prague 13 has to his name"

A look at the frequency of *na svém kontě* in the SYN8 corpus (a collection of a range of text types, 5,391,362,082 tokens) reveals that this pattern is considerably more frequent in journalism than in other texts (cf. table 21).

text type	frequency	i.p.m.
journalism (publicistika)	41,936	8.65
non-fiction (oborová literatura)	632	1.54
fiction (beletrie)	37	0.28
total	42,605	

Table 21. Frequency of *na svém kontě* by text type in SYN8

Another journalistic pattern is *moci se těšit na* "can look forward to", used to introduce or advertise an event (21), but also simply to introduce a future announcement or prediction (22).

21) od 20 hodin vystoupí skupina Booger z Mariánských Lázní . [...] Koho nepřiláká jen hudba, může se těšit na sele na rožni .
"the band Booger of Mariánské Lázně will be performing at 8 pm. [...] Whoever is not attracted by the music alone can look forward to pork roast"

22) V sobotu ještě potrvá nynější slunečné a teplé počasí. Můžeme se těšit na teploty kolem sedmnácti
"The current sunny and warm weather will continue till Saturday. We have temperatures around 17 to look forward to"

In SYN8, *moci se těšit na* was found to occur more frequently in newspapers than in other text types, compare table 22.

text type	frequency	i.p.m.
journalism (publicistika)	7,215	1.49
non-fiction (oborová literatura)	217	0.53
fiction (beletrie)	16	0.12
total	7,448	

Table 22. Frequency of *moci se těšit na* by text type in SYN8

In summary, the analysis of *na* patterns in context has revealed that the most frequent ones are multifunctional sequences of function words, often including punctuation and occurring around syntactic boundaries, similarly to many *v* patterns. Other *na* patterns are more specialised in the functions they perform: expressing numerical meanings (such as approximation: *na # tisíc*), referring readers to information outside the text (*na internetových stránkách*), subsuming the register-specific *na snímku* referring to a photograph accompanying the text. Other register-specific patterns *na svém kontě* and *moci se těšit na* were found to be more frequent in journalism than in other registers (according to the SYN8 corpus).

4.5.7 Czech prepositional patterns in newspapers: summary

The prepositional patterns identified in Czech newspaper texts were centred around a range of different prepositions – the most frequent one being *v*, apparently the most frequent Czech preposition overall. The individual analyses of selected prepositions and the patterning around them have offered a glimpse into the wealth of the different types of prepositional phraseological sequences, conveying a range of meanings and fulfilling various textual functions. Across different prepositions, one feature of distribution of patterns was shared – a good deal of recurrent patterning was found around syntactic boundaries, specifically introducing coordinate or subordinate clauses, often relative clauses. This points to a higher degree of conventionalisation in patterns which fulfil a specialised syntactic function.

Prepositional patterns were shown to carry a variety of meanings and functions which are to be expected in the newspaper register. Examples include: reference to past events or to events in progress around the present moment; representing personal agents and quoting them; or referring readers to extratextual information.

The findings have also confirmed the notion that journalism is not neutral informative discourse. This is reflected in a number of prepositional patterns manifesting semantic prosodies (cf. the unexpected negative evaluative prosody of *v souvislosti s*). It was also suggested that some prepositional patterns display semantic preference for particular contexts, delineated by the topic of the texts (sports) or the language employed in them (legal terms).

4.6 Prepositional patterns in English newspapers

In this chapter an analogous study to that on Czech newspapers is conducted using comparable English data, namely the SiBol (Siena-Bologna) corpus, comprising English journalistic texts, predominantly broadsheet newspapers, published between the late 1990s—2000s.[46] The focus is again on patterns containing a preposition.

4.6.1 Data & method

The corpus size is 768,941,480 tokens, which roughly corresponds to the Czech SYN2009PUB (844 881 368 tokens). For the purposes of the present study the corpus was restricted to texts from the UK and USA (henceforth subcorpus), yielding 630 million tokens (535 million words).[47] The SiBol corpus was accessed via SketchEngine.[48] First, lemmatised 3-5-grams containing selected (see below for details on the selection) prepositions were extracted from the custom UK/USA subcorpus of SiBol.[49] I identified the collocational patterns of prepositions to reveal their semantic prosodies (Partington, 2004). Left and right collocations were identified separately to pinpoint potential distinctive patterns bound either

[46]Anglophone newspapers published in 1993, 1995, 2010, 2013. A total of 1.5 million articles from 14 newspapers, mostly broadsheets.

[47] Newspapers from other countries were excluded in order to avoid employing data from countries where English is spoken mostly as a second language (i.e. India, Hong Kong, Nigeria and the Arab world), focussing only on the UK and US varieties.

[48] https://www.sketchengine.eu/
(Kilgarriff et al., 2014)

[49] Minimum frequency 5, nesting parameter on (i.e. overlapping n-grams are collapsed), lettercase disregarded.

to the element modified by the prepositional phrase, or in turn to the prepositional complement.

The Czech study focussed on the Czech preposition *v* (the most frequent preposition both in the corpus overall and in the n-grams) and *na* (second most frequent in the n-grams). Here I will be analysing two English prepositions, namely *of* and *in .* Arguably, *of* is one of the most frequent English prepositions (cf. Sinclair, 1991): it ranks as the second most frequent word in written English overall (Leech, Rayson & Wilson, 2001, p. 181; cited in Groom, 2010, p. 63). Likewise, in the *SiBol* corpus, *of* has a conspicuous presence: among the total 1000 n-grams retrieved, there are 215 n-grams containing *of.* Moreover, *of* occurs in 6 out of the top 20 n-grams retrieved.

Secondly I will examine the patters of *in*, to be compared with its Czech equivalent *v.* Admittedly, referring to translational equivalence in prepositions is inherently problematic: prepositions are highly frequent, while also highly polyfunctional and polysemic; and dictionary translations of prepositions tend to be rather simplistic – in practice, prepositions are translated in a variety of ways, certainly not restricted to direct translation equivalents (Klégr & Malá, 2009). Even with these limitations in mind, *in* seems a suitable option, since it is by far the most frequent translation of *v.* In the InterCorp corpus (Rosen et al., 2020), the Treq tool[50] (Škrabal & Vavřín, 2017; Vavřín & Rosen, 2015) retrieves 910,451 instances of *in* as the translation equivalent of *v*, followed by *at* with only 66,221 hits.

Within the extracted n-grams, an arbitrary cut-off point was placed at the frequency of 10,000 in both the Czech and English data, yielding the following results. The remarkably high frequencies of *of* patterns are in line with the previous findings about *of* being a top frequent English

[50] Treq version 2.1, available at https://treq.korpus.cz/index.php accessed 1 Sep 2021. Settings used: lemma, *v* from Czech to English, restricted to Collections.

preposition.

- *v* patterns: 167 n-gram types, 1,179,492 n-gram tokens, top frequent n-gram type comprised 204,506 n-gram tokens (henceforth as top frequency)
- *na* patterns: 50 n-gram types, 1,438,233 n-gram tokens, top frequency 165,272
- *of* patterns: 195 n-gram types, 4,269,442 n-gram tokens, top frequency 236,038
- *in* patterns: 75 n-gram types, 1,392,147 n-gram tokens, top frequency 64,338

The resulting n-grams were analysed with regard to their semantics and/or the textual function they fulfilled. The resulting inductively defined categories are presented in tables 1 and 2 in the next section.

Selected *of* and *in* patterns were subsequently analysed in terms of their left and right collocations[51] and, by extension, semantic prosodies. Left and right collocations are examined separately, with a collocation window of 3 positions on either side.[52] I aim to identify structural (parts of speech) as well as semantic patterning around the n-grams, paying attention to semantic preferences exhibited by individual patterns. The ultimate aim is thus to provide a complex overview of the prepositional patterns with regard to their relations to the broader context.

4.6.2 Pattern categories

In this section I provide a brief overview of the semantic/functional categories identified. Afterwards I will focus specifically on the functional

[51] By contrast, for Czech, the collocation analysis of *v* patterns focussed on right-hand collocations, since more collocations were found on the right-hand side.

[52] Positions are counted excluding the node, i.e. the node is 0, its preceding word is -1 etc.

category of complex prepositions, to be analysed in more detail.

4.6.2.1 *Of* patterns

The inductive categorisation of *of* patterns is outlined in table 23. The most frequent groups are discussed in more detail below.

category	n-gram tokens	n-gram types	example
quantification	1,730,941	60	per cent of, the rest of, a bit of
specification	510,835	35	the cost of, as a result of, a kind of, the likes of
place	390,839	23	of the country, in the middle of
title	332,203	16	chief executive of, a member of
time	324,864	16	at the age of, of the season
time/place	323,504	6	the end of the, of the game
preposition	253,349	12	in the wake of, in front of
attribution	201,648	15	the face of, of the royal, the author of
adjective	82,771	5	of the best, of a new, of the great
action/state nominalisations	49,534	4	the loss of, the sale of, the death of
conjunction	34,324	1	because of the
total	4,251,175	194	

Table 23. Semantic/functional categories of *of* patterns

The most numerous as well as formally varied group are *of* patterns expressing quantitative meanings. The most frequent quantitative pattern (as well as the most frequent pattern overall) is *one of the* (236,038 occurrences). Its strongest right-hand collocations (span 1-3, ordered by logDice to capture typical collocates) are superlatives (*most, biggest, best*) or adjectives whose meaning is close to superlatives in that they express an exclusive position (*leading, main*); the top collocates also include the nouns *world´s, reasons,* and *things*. The collocates suggest that *one of the* typically brings to the reader´s attention something significant or prominent.

The next biggest group are patterns comprising a noun phrase postmodified by the *of* phrase, serving to specify its meaning. The head of the noun phrase often has a relatively general, vague meaning (thus requiring specification), e.g. *the idea of, the use of, a kind of, a matter of*. Others refer to qualities such as *the cost/size /quality of*. Interestingly a number of the patterns refer to the worth of something (*the cost/value/price of*), which may correlate with the prominence of quantifying patterns in contexts referring to finances.

Patterns referring to place and time were also represented among the most frequent functional types. The most frequent place patterns refer to positions within a space (*the top/back/side/middle of*) – likely they are so frequent due to their meaning being broad and general. Other frequent place patterns referred to institutions: *the university/bank/house of* (the latter often referring to *the House of Lords* or *Commons*). This points out the types of institutions which are often discussed in the media, arguably in relation to news from the area of science or education, finance, or politics respectively.

Temporal patterns often included general patterns like *at the start/beginning of.* The prominent pattern *at the age of* proved interesting – its most typical left collocate is *died*, followed by related collocates *cancer, death, diagnosed* ; it also often occurs with *debut* and *retire(ment)*, referring to people´s careers. These collocations suggest it is often used in news about celebrities, in obituaries, or gossip columns etc.

The time/place patterns are those which could not be placed with either place or time as they can convey both meanings (*the end of the*).

The pattern group labelled "title" refers to job titles or similar labels attributed mostly to persons who are being either discussed or cited, and are introduced through their activity, membership in a group etc. Most refer to high-ranking, prestigious social positions such as *chief executive, head, director, chairman;* very frequent are variations on *a member of* (top right collocates are *committee, staff, Parliament, board* – again referring to people in positions of power and/or prestige).

Of also often forms part of complex prepositions such as *in front of* – these will be discussed in detail in the following section.

4.6.2.2 *In* patterns

Overall, the most frequent *in* patterns (in terms of n-gram token frequencies) were also represented by the most n-gram types, there was no salient mismatch in the token and type numbers. Their categorisation is presented in table 24.

category	n-gram tokens	n-gram types	example
place	376,009	17	in the world
quantification	261,849	12	in the first, per cent in
time	174,178	9	in recent years
preposition	162,408	10	in front of, in terms of
ambiguous[53]	96,427	7	in the new, up in the
copula	79,129	5	to be in
circumstances	59,806	5	in an interview, in an attempt to
other	182,341	10	was born in, in a way
total	1,392,147	75	

Table 24. Semantic/functional categories of *in* patterns

[53] Ambiguous = where no particular meaning or function was discernible in the pattern, or more than one meaning could be expressed.

Patterns with *in* most often express adverbial meanings of place or time (less often circumstances, cf. table 24). Temporal *in* patterns are varied as regards their meanings, ranging from relatively vague (*in the past, in the future*) as well as more specific (*in the morning, in the year*) temporal meanings.

Patterns referring to quantities and numbers were another frequent category. The patterns *in the first, last, final* help structure the text, describing the sequence of reported events. The pattern *for the first time in* was found to collocate with the right-hand collocates (span +1-+2, LogDice – by typicality) *history, decades, years, career, memory, ages*. It points out an important or unique event (cf. Malá et al., 2021). The items *in the first half, in the second half* either have temporal reference *(… of the year, of the 1980s*) or refer to sports matches *(, when Arsenal played beautifully*).

Patterns involving *in* which form a complex preposition were also frequent. They are especially relevant given that this study focusses on prepositional patterns, therefore they will be analysed more closely in the following section. Other categories included verbal patterns: involving a copula, or a verb followed by an *in*-phrase. *In* patterns did occur around syntactic boundaries (at the beginning of relative or co-ordinated clauses), but these instances were nowhere near as frequent or prominent as with the Czech prepositions *v* and *na*.

4.6.2.3 Complex prepositional patterns: analysis

Complex prepositions were identified among both *of* and *in* patterns. Remarkably, all but one of the *in* complex prepositional patterns were included in the *of* complex prepositional patterns (cf. table 25). For this reason, both groups will be described together.

of patterns	freq	*in* patterns	freq
out of the	78,319	in front of	34,733
in front of	34,733	in terms of	25,735
in terms of	25,735	in favour of	16,692
in favour of	16,692	in the face	14,626
on top of	14,951	in charge of	13,365
in charge of	13,365	in the case	12,085
out of a	12,959	in the wake	11,592
on behalf of	11,681	in the wake of	11,562
the wake of	11,591	in addition to	11,238
in the wake of	11,562	in the face of	10,780
ahead of the	10,981		
in the face of	10,780		
total	253,349	total	162,408

Table 25. Complex prepositions with *of* and *in* (patterns not containing both *of* and *in* are in bold)

The complex prepositions were examined through their collocations to reveal semantic preferences and/or evaluative prosodies. Left and right collocates were examined separately for each selected prepositional pattern. Collocations were ranked by the LogDice metric in SketchEngine in order to prioritise their typicality.

The following sections summarise the cases where semantic preferences or prosodies were identified. The most frequent complex prepositional pattern was *out of the.* Its typical left collocates were mostly motion verbs, implying either passive or forced motion (*pulled, knocked, thrown, dropped, forced*) or active motion (*coming, moved, walked, stepped*). Right collocates referred to places (*window, door, house, room, closet, box*) or events (*race, game*). A number of right collocates were instances of idiomatic, metaphorical sequences: *blue, question, woods, ordinary, equation*. The pattern *out of the box* may be either literal (23) or metaphorical (24) in meaning.

23) a computer that comes ready-to-run out of the box

24) [...] George Bush's presidency to try to break out of the box within which the United States and the United Nations have confined him

In comparison, the pattern *out of a* is less frequent (*out of the* had 78,319 hits; *out of a* only 12,959). It shares some collocate types with its variant *out of the*: on the left, verbs of passive (*pull, thrown, dig*) and active motion (*walk, step, jump, storm, grow)*; on the right, words referring to locations (*window, hole, bunker, suitcase, hat*). However, *out of a* has some extra right collocates: *out of a total* describes numerical data, *out of a desire* describes motivations for actions, and *out of a molehill* points to the idiom *make a mountain out of a molehill*.

The pattern *in front of* illustrates that prominent left collocates do not necessarily occur together with the right ones, and each group of collocates may be associated with different text-types. Its left collocates are verbs indicating position (*sit, stand, parked, gathered*) and the deverbal nouns *protest, demonstration,* referring to events where the activity reported by the verb occurred. These uses of *in front of* are likely to occur in reports of current events. By contrast, right collocates come from markedly different contexts: entertainment (*crowd, audience, fans, camera*) or sports (*goal*). Other typical collocates were *mirror* and *television*, which can refer to a wide range of people (unlike e.g. *in front of the audience/goal*, restricted to public figures).

In favour of shows a semantic preference for legal or political contexts: left collocates include words related to decisions of political entities or courts (*vote*, *ruled*), the more generic *argument*, as well as clearly evaluative collocates pointing to an imbalance (*weighted, biased, skewed*): these imply that though *in favour of* itself expresses positive meaning, its evaluative prosody seems rather negative. Right collocates were varied: the subjects of the voting, ruling etc. (*motion, ban, strike,* *amendment)*, the surprising *marriage*,[54] as well as the apparently contradictory *retaining, keeping* as opposed to *reform* and *action.*

Similarly, *ahead of the* has a preference for contexts related to statistics (*curve*) and political events (*elections, poll, assembly, summit*). An idiomatic pattern is *ahead of the pack*.

Finally, two *of* patterns exhibited distinct evaluative prosodies. In case of *in the wake of*, some left collocates are neutral (*come, introduces, assumes, significance*), others suggest negative contexts (*gun; resign, quit*). The negative prosody is fully revealed by the right collocates: the majority of the top 20 refer to negative phenomena (*scandal, crisis, bombings, hurricane, shootings, collapse, tragedy,* including the topical *Katrina, Sandy* and *LIBOR*[55]).

In the face of typically occurs as part of the idiomatic *fly in the face of*. Apart from that, it is preceded by adjectives and nouns referring either to positions of strength (*resilience, courage, bravery, dignity*) or the contrary (*powerless, helpless*). Typical right collocates refer to challenging (*competition, odds*) or unfavourable conditions (*adversity, opposition, criticism, hostility, onslaught, provocation, threats*), which may be emphasised by a premodifier (*mounting, overwhelming, stiff, fierce*). The left and right collocates are clearly related, combining to portray an agent coping more or less successfully with a difficult situation.

[54] The sequence *in favour of marriage* was marked as a typical collocate through the LogDice metric; admittedly it only occurred 16 times – it seems to typically refer to marriage generically, as a public institution or a political topic (cf. *the Church had previously argued strongly in favour of marriage*).

[55] The acronym LIBOR stands for London Interbank Offered Rate. During the time period covered by the corpus, newspapers reported on the LIBOR scandal, cf. https://www.bbc.co.uk/news/business-19203103.

4.6.3 English prepositional patterns in newspapers: summary

Both *of* and *in* patterns often express quantitative meanings, as seen in the top frequent pattern overall, *one of the* (which collocates with superlatives and serves as a means of highlighting). Similarly, *for the first time in* collocates with *history, decades* etc., bringing a unique event to the reader´s attention. Other *of* patterns often contain generic abstract nouns (*the idea of, a kind of*) whose meaning is specified by the *of* phrase. The collocates *cost, value, price* seem connected with the frequent occurrence of quantitative patterns.

Topics or entities discussed in the press were revealed by temporal and spatial adverbial patterns (institutions *the university/bank/house of*, lives of public figures *death/retirement at the age of*) and patterns referring to people´s (mostly prestigious) job titles (*chief executive of*). Other patterns reflected the need for precise temporal descriptions of events or their sequences (*in the first half*).

Complex prepositions including *of* and *in* largely overlapped. The most frequent complex preposition *out of the* collocated with verbs of motion on the left (*pulled, knocked, coming, walked*); and locations on the right (*window, room*). Besides, it occurred in idiomatic patterns (*out of the blue/equation*). The left collocates of *in front of* (*sit, stand, protest, demonstration* – locations of events) occurred in different contexts than its right collocates (*audience, goal* – entertainment and sports). *In favour of* exhibited a semantic preference for legal and political contexts, as well as a slight tendency towards negative evaluation (*biased*), in contrast with the positive lexical meaning of the pattern itself. Finally, *in the wake of* and *in the face of* have shown negative evaluative prosodies (e.g. *in the wake of* + *crisis*; *bravery/dignity + in the face of* + *adversity/onslaught*).

4.7 Conclusions

4.7.1 Newspaper phraseology: summary of results

This study examined the register of newspaper texts through the lens of n-grams. First, a pilot study was conducted on English data, extracting n-grams, observing what features of the texts this method revealed and what methodological problems arose from it. The n-gram search pointed to the importance of direct speech (quotations), and aboutness words suggestive of prominent topics (current affairs, sports, politics, institutions). The results suggested that a focus on grammatical words in patterns may be revealing: grammatical patterns highlighted an imbalance in the representation of male and female speakers quoted by the newspapers. Including punctuation in the n-gram search showed that a number of recurrent patterns contained punctuation marks, i.e. they were found around syntactic boundaries, such as introducing direct speech.

Crucially, a major effect of corpus design was found – references to sports, a topic frequently discussed in tabloids, were overrepresented in the n-grams. This highlights the importance of bearing in mind the text-type composition of the corpus when interpreting the patterns identified. As suggested in the pilot study, the notable frequency of sports-related patterns may be caused by the fact that the language of sports reporting is highly repetitive, allowing fairly little space for language creativity.[56] Overall, the n-grams yielded interesting but fairly isolated findings, hence in the follow-up study attention was also paid to frequent contexts of the patterns, using collocations.

The following step was a contrastive study of English and Czech newspapers. The scope was narrowed to prepositional patterns, since prepositions (due to their frequency and even dispersion) allow for the

[56] On the other hand such creativity does seem to be introduced by Czech journalists through stylistically marked expressions, as seen in ex. 16 (*shrine* meaning goal).

identification of a variety of frequent and pervasive recurrent patterns. Two frequent prepositions were selected for each language: *v, na, of, in.* The n-gram method was complemented by an analysis of left and right collocations of the selected prepositional patterns, aiming to reveal semantic prosodies or preferences.

The results suggest that in both Czech and English, prepositional patterns can convey a range of meanings which correspond to the informational function of the newspaper register, such as: reference to past events or to events in progress around the present moment; temporal reference and specifying the sequence of events; quantification; representing personal agents and quoting them; or referring readers to extratextual sources of information (e.g. website links). The findings have also confirmed that newspaper texts are not purely informational, as reflected in prepositional patterns manifesting particular semantic prosodies (e.g. the negative evaluative prosody of *v souvislosti s* or *in favour of*). This suggests that the reported events and entities are described in an evaluative light. However, the analysis of evaluative prosody raises a more general question of the delineation of evaluativeness. Even though some prepositional patterns manifest evaluative prosodies when occurring in newspapers, this alone does not answer the question whether (and how) such evaluativeness contributes to newspaper texts carrying any specific bias. This question needs to be addressed with the help of a close reading of particular texts – the methods of (critical) discourse analysis may be of assistance here. Further, some patterns serve the function of highlighting specific information, e.g. *one of the* combined with superlatives. Some patterns display semantic preferences, given by the topic of the texts (e.g. *ahead of the* seems associated with current affairs, *in the first half* or *v # minutě* with sports) or the language employed in them (e.g. *in favour of, v souladu s* combine with legal terms).

4.7.2 Cross-linguistic comparison

From the contrastive perspective, the results allow for a comparison of patterns containing the translation equivalents *v* and *in*. Both patterns occurred with a similar frequency (1,392,147 instances of *in* patterns, 1,179,492 *v* patterns) but the Czech *v* patterns are more formally varied, comprising more than twice as many n-gram types (167 *v* n-gram types as opposed to 75 *in* ones).

In both languages, adverbial patterns were among the most frequent functional groups, especially those referring to place, time or circumstances. The Czech temporal patterns often referred to the time of a particular planned event (*dnes v ___ hodin 'today at __ o'clock')* while the English ones tended to convey broader temporal meanings *in the past, future, recent years, in the year*.

Both languages shared the group of patterns expressing quantitative meanings or precise temporal reference (*in the first*, *second, last*). The equivalent pattern *in one of the / v JEDEN z* may be followed by a superlative, pointing out salient phenomena.

Patterns comprising verbs were more frequent in Czech, where they often refer to past and future events (*se uskutečnit v 'take place in'*), while the English equivalent group only comprised the verbs *said, born, involved*, referring rather to persons.

Together with the adverbial temporal patterns, these findings suggest that Czech newspapers more often than English employ *v* patterns to inform readers e.g. about cultural or sports events.

4.7.3 Methodological findings

The methodological aspects which were added to the n-gram method in case study II (as opposed to the previous case study I) were mainly the inclusion of punctuation in n-grams, working with lemmatised data, and combining n-grams with collocation. In this section I will briefly evaluate these additions to the method and discuss some other areas which still remain to be explored.

Lemmatisation was introduced to allow for collapsing n-grams which only differ in inflection (prototypically in morphological suffixes). This was seen as advantageous especially in Czech. However, some results confirmed that lemmatisation can be problematic, concealing relevant differences between patterns (Čermáková & Chlumská, 2016; Granger, 2014). This was shown e.g. on the lemma-gram V TENTO DEN, whose plural form *v těchto dnech* ("in these days") referred to current issues in progress, while its singular variant *v tento den* pinpoints a particular date.

Further, in some patterns lemmatisation turned out to have no major advantage, as those patterns were found to occur in invariable forms, the lemma-n-grams being identical with the word-form n-grams, as was observed in complex prepositions. Hence, the overall contribution of lemmatisation is questionable. To reliably assess to what degree lemmatisation is beneficial would require a focussed comparison of lemmatised and non-lemmatised n-grams retrieved from the same dataset, which is beyond the scope of the present study. Still, lemmatisation will be employed in the following case study III, in order to test lemmatisation on another pair of corpora representing a functionally very different register – children´s literature.

Allowing for punctuation in n-grams was another parameter which was newly introduced in case study II. As early as the pilot study, it was found that some n-grams only differed in punctuation, (as in *. He said :* and *He said :"*). Had punctuation been excluded, some of those n-gram hits would have merged into one. This was one reason for including punctuation in n-grams in the case study proper. Yet this same example also points towards a potential caveat: punctuation is contained at the expense of wordforms or lemmata, hence the resulting n-grams reflect less about the lexical meanings of patterns. To relate this to the abovementioned example, excluding punctuation, we may have learned more about what words commonly precede or follow the sequence *he said*. Further, due to the frequency of punctuation and overlaps between n-grams, a number of patterns may be represented multiple times, as was pointed out in the pilot study. On the other hand, punctuation in n-grams can be valuable in pinpointing how phraseological sequences tend to form around syntactic boundaries, how they contribute to structuring texts (e.g. reflecting common ways of linking super- and subordinate clauses). Also, it pointed to particular uses of punctuation being frequent and fulfiling important discourse functions relevant to the register at hand – specifically, introducing direct speech. To sum up, my results suggest that including punctuation has both its advantages and disadvantages – it may yield interesting results, but it poses some limitations on the results, as less will be found about the lexical meanings of patterns.

Overall, the study has pointed towards several aspects of the newspaper register that can be efficiently revealed by n-grams: namely complex prepositions (which proved a fruitful starting point towards identifying semantic prosodies and preferences through collocations), or lexical style markers which proved typical of the newspaper register in comparison with a general reference corpus. Hence the n-gram method seems suitable for identifying register-specific phraseological elements. However, the results still indicate that complementing n-grams with another method is advisable since it provides a more detailed, comprehensive and revealing portrayal of the register phraseology. In this study, n-grams were combined firstly with a predetermined item, specifically a grammatical word - preposition, in order to identify patterns which are evenly dispersed throughout the data and contribute to text-structuring. As a next step, collocations of the identified patterns were explored to reveal how the patterns are employed in context and contribute to textual meanings, including evaluative ones. N-grams containing a function word were found to be an efficient gateway towards patterns with text-organising functions. To complement this study, testing n-grams based around selected lexical words seems a logical next step. This approach will therefore be employed in the following case study III.

5 Case study III: an n-gram analysis of children´s fiction[57]

5.1 Introduction

The previous two chapters introduced the results of two case studies – corpus-driven n-gram analyses of Czech and English parliamentary debates and newspapers – which have identified several methodological problems. Essentially, the difficulties pointed towards the need for several parameters to be considered. Specifically, these were:

- Morphological variability: can be resolved by lemmatisation (as employed in case study II).
- Word-order variability is resolved by disregarding the order of words/lemmata within an n-gram, resulting in 'shuffled' n-grams (to be employed in case study III).
- Overlapping n-grams: a shorter gram may be contained within a longer one, as seen in case study I; this will be addressed in case study III by collapsing n-grams differing in one position.
- Lexical variability: to capture the lexical variability of patterns, we may allow empty slots within n-grams, resulting in 'skipgrams' (Cheng et al., 2006). This approach will likewise be employed in case study III, through collapsing n-grams differing in one slot.

A combination of these steps may help to accommodate for the

[57] Hereby I acknowledge that the results of research discussed in this chapter had previously been presented and published as follows:
- preliminary results at a work-in-progress stage (Šebestová & Malá, 2019);
- paper presented 2019 at ICAME conference (Šebestová et al., 2019);
- final results summarised in (Malá et al., 2021, written 2019, published 2021);
- the results of this study were also employed as the basis for further research in (Malá, 2019).

This chapter has been adapted from the co-authored Malá et al. (2021) – formal and wording changes have been made but the results are identical. Verbatim sections are indicated. I declare that I have the co-authors´permission to present the results within this dissertation.

typological characteristics of both Czech and English, although admittedly each of these parameter adjustments may also entail its drawbacks. The present chapter adopts a synthetic approach, addressing these previously encountered problems through adjustments made to methodology.

5.2 Motivation

5.2.1 Register characteristics of children's fiction

In the present case study I focus on fiction intended for young readers, aiming to explore its phraseological characteristics. As a register, children's fiction is defined on the basis of its intended audience. This determines its aims and functions: first and foremost, it is presumed to play an important role in a child's cognitive development as well as promoting socialisation, acquainting the reader with a range of social and cultural norms. It represents the young reader's first point of contact with literary language (Čermáková & Chlumská, 2016, pp. 162–163), and more generally with various uses of language. The young reader is prompted to "compare their experiences with the experiences of others" (Stephens, 2005, pp. 73–74). Thus, children´s fiction is perhaps best characterised by having a host of functions to fulfil: as Hunt points out, "children's books are used for different purposes at different times – for more things than most books are" (Hunt, 2005, p. 10). The abovementioned objectives of children's fiction shape its specific linguistic properties as well as its content. In the following section, drawing on selected previous researches, I will outline some observations about the nature of children´s fiction and its language, suggesting why it is a potentially rewarding area for linguistic inquiry.

5.2.2 Simplicity vs. complexity

Linguistic characterisations of children´s fiction seem fraught with a paradox: it is described in rather contradictory terms. The language of children's literature is often perceived as – or, by extension, expected to be – simplified in comparison with adult fiction language, "a 'scaled-down' version of 'language in general'" (Thompson & Sealey, 2007, p. 2). There is a widespread presumption that the authors of fiction for children modify their language to facilitate understanding, such modifications being dictated by the authors´ preconceptions about the needs and abilities of their intended readers (Hunt, 2005, p. 73; Nikolajeva, 2005, p. xv). Apart from linguistic simplification, children´s fiction is viewed as simple in terms of its narrative structure (e.g. straightforward plot and characters, chronological order of events) and discourse (clear viewpoint, omniscient narrator). Other features stereotypically attributed to children´s fiction include a bias towards positivity and happy endings, favouring action rather than psychological depth, a predominantly didactic focus, and repetitiveness (Nikolajeva, 2005, p. xiii-xiv).

The notion of children´s literature being simplified is corroborated by some of Thompson and Sealey´s corpus study findings (2007). While children´s and adults fiction was found to contain similar core vocabulary, a closer lexical analysis suggested that "fiction written for children may make greater use of the more literal meanings of words and less use of their more figurative or metaphorical meanings than the discourse found in the comparison corpora" (of adult fiction and newspapers respectively) (ibid., p. 16). This is illustrated on selected words for body parts, which were found to be employed in less abstract contexts in children´s fiction – likely resulting from the writers´ presumption that children learn to understand abstract senses later on (ibid., p. 18).

The assumption that children´s fiction tends to be linguistically straightforward or even pared down is however challenged by its enormous functional load and socio-cultural significance as outlined earlier. Hunt argues in favour of children´s books being remarkably complex, going so far as to claim that "we are dealing here with fundamental questions of communication and understanding between adults and children" (Hunt, 2005, p. 2). Notably, children´s literature is characterised by an asymmetry in the writer-reader relationship: the adult writer´s experience, both linguistic and otherwise, is fundamentally different from that of the child reader´s (Nikolajeva, 2005, p. xv). Thus, authors face an immense “challenge of writing for children, and about children’s concerns”, trying to adopt the implied young reader´s mindset and perspective (Thompson & Sealey, 2007, p. 4). The issue is further complicated as our understanding and views of childhood are constantly evolving, conditioned culturally as well as historically (ibid.: 3).

Thus, despite its perceived simplicity, research on children’s literature has indicated that the relationships between characters and their environment are specific in children’s fiction when compared to adults’ (Hunt, 2005, p. 3; Thompson & Sealey, 2007, p. 4). To summarise:

> Children’s books are different from adults’ books: they are written for a different audience, with different skills, different needs, and different ways of reading; equally, children experience texts in ways which are often unknowable, but which many of us strongly suspect to be very rich and complex. (Hunt 2005: p. 3)

The seemingly contradictory nature and specific status of fiction for young readers gives rise to the following overarching research questions: What characterises the language of children´s fiction? Does it display features such as simplification or repetitiveness? Therefore, children´s fiction invites examination of its phraseological properties. Employing phraseological sequences as a unit of analysis seems productive since they are register-specific (e.g. Biber et al., 2004), to the point of being able to function as register signals (Hyland, 2008, p. 5). Studying phraseologies in children´s literature can provide valuable insights because recurrent lexico-grammatical patterns are likely to be stored in the readers´ mental lexicon as representative of typical usage, considering that the reader is exposed to them repeatedly and at an early age (Čermáková & Chlumská, 2016, p. 164). Further, the cross-linguistic perspective may reveal some culturally distinctive features of this particular register. Cultural factors are likely to influence how register shapes phraseology. Phraseology has been shown to be closely linked with culture, as testified by idiomatic expressions drawing on references to phenomena associated with a particular community (Sabban, 2008). Cultural norms may translate into the phraseology of registers, especially those which have a specific cultural load, such as children's literature. Yet, disentangling culture-specific links from more universal ones is extremely complex (Colson, 2008). Hence, I will pay attention to linguistic differences potentially reflecting the different cultural landscapes associated with the languages compared; however, the primary focus of my study remains on identifying and comparing salient phraseologies in the two corpora at hand.

5.3 Time in children´s literature

Due to the necessary limitations posed by the scope of this research, this case study only examines a corpus of fiction intended for children and teenagers, not comparing it to fiction for adults. Further, the cross-linguistic phraseological analysis presented in this chapter focusses on a selected semantic area, namely the expression of time. Time has been shown to be framed and represented in idiosyncratic ways in children´s fiction (Thompson & Sealey, 2007). According to Nikolajeva, actions and events are often repeated in children´s literature, possibly "since the iterative reflects a child's perception of time as cyclical, non-linear" (2004, p. 167). Yet Nikolajeva also offers a conflicting account, stressing the importance of linearity in the Western tradition of children´s literature (2000: 5, cited in Sainsbury, 2014, p. 189). Sainsbury relates this to the claim that "[narrative plays a] central role in making sense of the shared human experience of being in time." (2014, pp. 189–190).

Representing the concept of time seems to contribute to the socialisation mission of children´s fiction, as "the acquisition of time sense is part of the socialization process, working alongside the learning of language skills and moral values." (ibid., pp. 187–188) Moreover, as Pinsent argues, "specific forms of chronotope (i.e. the indivisible 'unity of time and space' in a work of literature) are unique for particular genres." (2014, p. 109) The significance of time and space in fiction overall is further supported by Thompson and Sealey´s study, revealing a majority of nouns in (both adults´ and children´s) fiction corpora to carry spatial and temporal meanings (2007, p. 14). Hence, my analysis will be centred on the representation of time. In turn, the expression of spatial meanings in Czech and English children´s fiction has been examined by Čermáková and Chlumská (2016, 2017), allowing for potential comparison.

I focus on temporal meanings expressed by adverbials, not by the grammatical categories of the finite verb (tense and aspect). Temporal adverbials can be constituted by noun phrases (e.g. *tento večer*, *this evening*), prepositional phrases (e.g. *o Vánocích*, *at Christmas*), adverb phrases (e.g. *nakonec*, *finally*), or clauses (e.g. *až přijedete*, *when you arrive*). The temporal meaning is signalled by the head noun (in noun and prepositional phrases), adverb, or by a temporal conjunction. Though temporal relations may also be expressed by prepositions, due to the vagueness and ambiguity of their lexical meaning they cannot serve as reliable markers of temporal meanings (cf. Hasselgård, 2017, p. 82). In addition to adverbials, temporal meanings can be conveyed by noun phrases with a temporal head noun which function as nominal clause elements, e.g. the subject in *<u>Our time</u> has come*.[58] [59]

5.4 Children´s fiction in Czech and English

Czech-English contrastive phraseological studies employing n-gram methodology have been relatively scarce to date, with the notable exception of Čermáková and Chlumská´s pioneering papers (2016, 2017). They compared Czech and English children´s fiction, using n-grams to

[58] Examples come from *BNC-Jun* and *SYN-7-Jun* (see section 3 for details).

[59] As will become apparent in the analytical section later, these syntactic functions, however, do not appear to be performed by patterns based on frequent n-grams explored in this study.

identify expressions relating to space. Their comparison showed that space n-grams were the least frequent in Czech originals, more frequent in translations into Czech, and the most represented in English originals. The results indicated that while the English texts favoured a more specific and/or exact representation of space, in Czech the spatial references were more vague and less prevalent overall. The Czech translation equivalents tended to be more condensed in nature, including omissions, single-word prepositions as equivalents of English complex ones, or containing verbal prefixes (e.g. *he was coming to the top of the steps – docházel na vrchol schodiště*). Overall, their analyses suggested that the two languages differ in the degree to which they employ the idiom principle (Sinclair, 1991), and in the degree of variability this entails.

Based on these findings, I expect the temporal patterns identified in my data to display differences between Czech and English in terms of their phraseological properties. Čermáková and Chlumská´s results point out that the length of equivalent n-grams may not correspond between Czech and English, given the fact that Czech may favour more condensed means of expression such as simple prepositions or prefixation, or even omission (2017). This is one argument in favour of extracting several different lengths of n-grams, which was likewise suggested by case study I, this volume. (The method will be discussed in more detail in section 3.) I further expect Czech temporal patterns to display a greater degree of morphological variability than English ones, reflecting the languages´ typological properties: while English is analytical, Czech has a complex system of inflections. Likewise, more lexical variability within Czech patterns is to be expected. Čermáková and Chlumská (2017, p. 83) have observed that their Czech fiction corpus yielded a wider variety of n-grams, while the English counterparts were more repetitive and uniform. This was attributed to Czech being morphologically variable as well as having flexible word order.

5.5 Material

The material was sourced from two subcorpora of the *British National Corpus*[60] and the Czech *SYN*-7 – cf. table 26[61]. Each of these general representative national corpora was restricted to lemmatised, original fiction intended for a child and/or teenager audience. Whereas the *BNC* contains text samples, *SYN-7* is composed of full texts; hence the difference in number of texts.[62] The two corpora do not entirely correspond in terms of publication time spans of the texts: BNC contains fiction published between 1960 and 1994 (75% after 1975)[63], *SYN-7* spans 1967—2014.[64] I have opted for these corpora nevertheless since they are extensive, balanced and representative; and to my knowledge, no other English corpora of more recently published children´s fiction were freely available.

language	English	Czech
corpus	BNC-jun	SYN-7-jun
= subcorpus of	*BNC* (1994)	*SYN-7* (2010)
size - tokens	2 046 755	2 821 044
size - texts	76 (excerpts)	59 (whole books)
text publication dates	1960-1994	1967-2013
no. of authors	36 + 29 adapted classics[65]	43

Table 26. The corpora used

[60] http://bncweb.lancs.ac.uk/

[61] https://kontext.korpus.cz/first_form?corpname=syn_v7

[62] The category of texts intended for a child and teenage audience was limited to fiction (i.e. excluding academic prose, non-academic prose and biography, and other un/published written material).

[63] *BNC User Reference Guide* http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#BNCcompo [accessed 30 July 2020]

[64] *Corpus SYN version 7* https://wiki.korpus.cz/doku.php/en:cnk:syn:verze7 [accessed 30 July 2020]

[65] Oxford Bookworms edition – adapted texts.

I expect to find patterns conveying functions characteristic of fiction for young readers. The scope of my analysis is determined by the overarching perspective connecting the three case studies presented in this thesis: i.e. I am predominantly concerned with phraseological characteristics, not aspiring to provide a comprehensive description of each respective specific register. Admittedly, the research question posed in the present case study would warrant the use of reference corpora of fiction oriented at an adult audience, in order to identify the salient features of children´s literature in contrast to adults´. However, a keyword analysis is beyond the scope of this thesis. Therefore, I do not conduct a keyness study at this stage, merely recording it as a potential next step for a follow-up research to complement the findings of this phraseological analysis.

5.6 Method[66]

The method employed here falls under the approach termed "from lexis to n-grams" by Stubbs (2007b, cited in Lindquist & Levin, 2008, p. 145): combining n-gram methodology with a lexical point of departure. This seems advantageous since n-grams alone are often neither structurally complete nor meaningful (as shown e.g. by Biber et al., 1999). Since my aim is to explore how the concept of time is commonly expressed in children's narrative fiction, I extract n-grams containing a temporal node: first, frequency lemma lists are extracted and examined for particular lexical items, which are subsequently used as the temporal nodes of n-grams. The resulting n-grams then serve as an initial quantitative step, aimed towards the identification of patterns and their subsequent qualitative analysis.

As the study focusses on the expression of time, the n-gram nodes are identified based on a corresponding semantic criterion. Relevant n-gram nodes were found based on frequency lists of the 500 top frequent

[66] The sections *Method*, *Analysis and Conclusions* are taken largely verbatim from Malá et al., (2021).

lemmata in the respective Czech and English corpora. I searched these lists manually for all words carrying temporal meaning. The manual search was performed separately by two annotators and then compared in order to control for omissions.[67] The sub-lists were then sorted by part-of-speech and compared between the two languages. The parts-of-speech represented were nouns, adverbs and conjunctions. The resulting lemma lists are summarised in table 27. These lemmata were then used as nodes in any given position in 3- and 4-grams, allowing for positional mobility within the n-grams. Some node lemmata were represented by translation equivalents in both languages, as in *when – když, moment – chvíle,…* - these are marked in boldface in table 27.

Lemmata which may refer both to time and another semantic field were excluded, such as *end*, which occurs in temporal as well as spatial expressions. Admittedly, this narrows down the scope of the study and makes direct comparison with the study by Čermáková and Chlumská (2016) more complicated; however, it helps avoid a considerable degree of semantic ambiguity.

	top-frequent temporal node lemmata
English (BNC-jun)	then, when, now, time, again, day, never, night, year, once, while, ever, always, moment, suddenly, soon, until, morning, minute, later, hour, week, already, evening, sometimes, late, ago, often
Czech (SYN7-jun)	už, když, pak, teď, den, chvíle, hned, brzy, rok, nikdy, pořád, čas, potom, znovu, noc, jednou, dlouho, najednou, ráno, hodina, doba, dnes, kdy, večer, konečně, stále, vždycky, nakonec, již, teprve, někdy, zatím, zítra

Table 27. The most frequent time-related lemmata[68]

[67] Thanks to Markéta Malá for her help with this.

[68] Table 27 is taken verbatim from Malá et al. (2021).

Next, I searched for n-grams containing each node lemma using the custom-developed *Engrammer* software (Milička, 2019). This tool has been specifically designed to be used with data in typologically different languages: importantly, it allows for extracting lemmatised and unordered n-grams (i.e. allowing for positional variability within n-grams, and merging n-grams differing only in word/lemma order). Both comprising a temporal lemma (rather than a word-form) and allowing for a variable position of the lemma in the n-gram expand the range of temporal multi-word expressions identified in the corpora, and facilitate English-Czech comparisons. N-grams are ranked by the strength of their association with the selected word form/lemma used as the n-gram node. The association metric employed here was the lower limit of the Risk Ratio confidence interval.

For both languages, n-grams were defined as "recurring strings, with or without linguistic integrity" (Lindquist & Levin, 2008, p. 144). I extracted 3-grams and 4-grams comprising the temporal node lemma at any position within the n-gram. The lemma *DAY*, for instance, can occur in the singular or plural form, and be placed in the initial, medial or final position within the n-gram (ex. 1a). The advantages of searching for n-grams comprising a specific mobile lemma are even more readily observable in Czech. For example, the lemma *DEN* ('day') may vary in number, case and position within the 3/4-grams (ex. 1b), and 3-grams which differ merely in word-order can be collapsed (ex. 1c).

(1) a. day and night, one day when, in those days
b. *den a noc* ("day and night "), *tři dny a tři* ("three days and three "), *ve stejný den* ("on the same day ")
c. *jednoho dne se* ("one day *se*-reflexive "), *se jednoho dne* ("*se*-reflexive one day ")

The choice of lemmata over word-forms is motivated by the effort to identify as many patterns involving the selected temporal node word as possible. Without lemmatisation, some low-frequency patterns would likely go unnoticed – particularly in Czech, due to its morphological variability. Yet a closer examination of the patterns reveals that typically, a pattern with a specific function contains only a particular word-form (rather than the full lemma). This is true for both English and Czech patterns (e.g. *day by day* rather than **days by days; hodnou chvíli -* singular "a nice moment " rather than **hodné* chvíle - plural "nice moments "[69]). This finding supports the notion that our identified chunks are indeed patterns, phraseological units, rather than free word combinations, as they are (to a degree) fixed. This is in line with Sinclair´s (2004, p. 31) observation that the collocates of a singular word form may not overlap with those of its plural variant (cf. Sinclair´s examples: *blue eyes* rather than *blue eye; in your mind´s eye* not **in your mind´s eyes*).

The length of the n-grams explored in this study, i.e. 3-4 words, is given by its semantic focus on the expression of temporal meanings: while 2-grams are usually difficult to classify semantically since they comprise predominantly grammatical words[70], 5-grams and longer multi-word units are not only less frequent but also more likely to be semantically complex and polyfunctional, potentially causing problems were they to be analysed together with shorter and mostly monofunctional units.

Typically, n-gram studies exclude punctuation from the n-gram search. This is in line with the related notion of collocation – usually collocations are conceived as not extending across syntactic boundaries (cf. Nesselhauf, 2004). However, since in Czech subordinating conjunctions

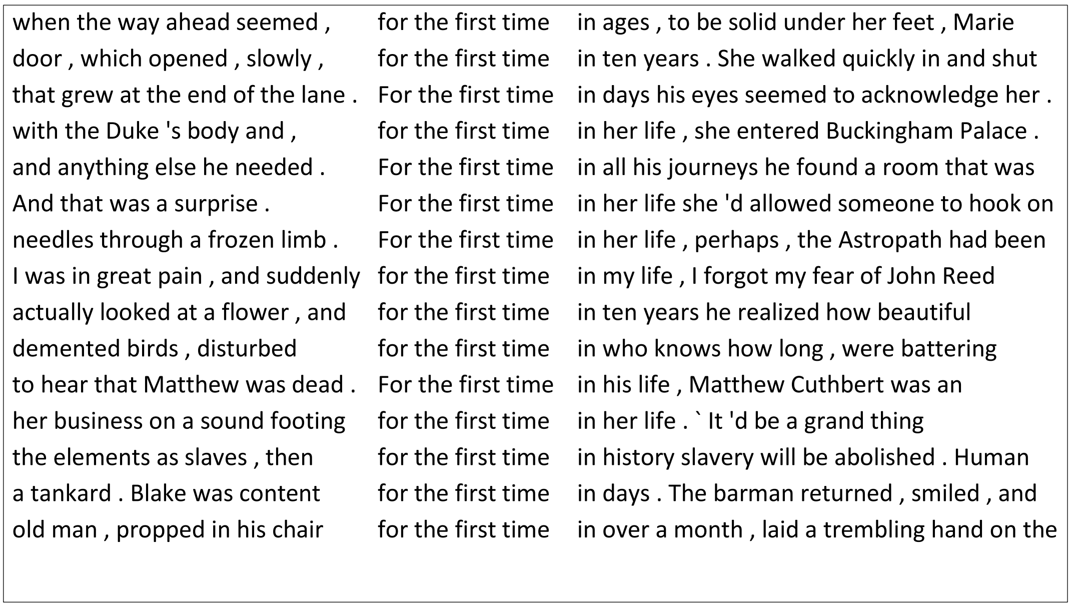
[69] Interestingly, in case of *hezkou chvíli* ('a pretty long time', lit. 'a nice moment'), replacing the singular word form with plural would change the meaning of the noun phrase and result in a loss of idiomaticity (the plural *hezké chvíle* has the literal meaning of 'nice moments' and cannot be used as a time adverbial).

[70] Moreover, "shorter bundles are often incorporated into more than one longer lexical bundle" (Biber et al., 1999, p. 990). This implies that a number of bigrams will often be included within longer grams extracted from the same corpus. Cf. also the results of case study I.

are obligatorily preceded by a comma, as in *chvíle, kdy jsme* (“moment when we “), punctuation was included in the n-gram search, because n-grams containing commas seem to capture more realistically the form of the temporal expressions in Czech. The results also contained n-grams with quotation marks, highlighting the ample representation of direct speech in children's fiction in both languages and its role in structuring the narrative, e.g. *'...Somebody should stop him!' Just then Martha ran into the room*.

As a next step, on the basis of frequent n-grams, recurrent temporal patterns were identified in both languages. Following Lindquist and Levin (2008, p. 144), I use the term “pattern” for “meaningful, linguistically structured recurring sequences of words”. A qualitative analysis of the patterns in context revealed the textual functions of the patterns in each language. This identification of patterns was assisted by the *Engrammer* functionality of merging similar n-grams, defined as grams differing in 1 slot, as in *for the first time + for the first moment*, or having 1 extra slot (*for the first + for the first time*). Conveniently, the variable slots can be displayed upon clicking the merged n-gram, which enables the examination of lexical variability associated with a selected pattern, as well as the identification of schematic patterns involving an open or variable slot.

To illustrate this, let us observe the lemma *TIME:* it frequently occurs (125 times) within the 4-gram *for the first time*. This 4-gram turns out to be strongly associated not only with *TIME* (out of its 144 occurrences, only 19 comprise an expression other than *time*), but also with expressing temporal meanings overall: *TIME* alternates in its slot mostly with other temporal noun phrases, e.g. *couple of weeks*, *few years*, *six months*, *hour*, *week*, *night*, *two days*, etc. This reveals that there is the phraseological sequence *for the first* NP*,* which displays a semantic preference for words denoting units of time, and therefore is typically employed to express temporal meanings. Upon expanding the context in which the n-gram *for* *the first time* occurs, the concordance lines (cf. Fig. 1) reveal a temporal pattern "*for the first time in* 'a long period of time'", associated with intensification and with signposting a turning point in the narrative.



when the way ahead seemed ,	for the first time	in ages , to be solid under her feet , Marie
door , which opened , slowly ,	for the first time	in ten years . She walked quickly in and shut
that grew at the end of the lane .	For the first time	in days his eyes seemed to acknowledge her .
with the Duke 's body and ,	for the first time	in her life , she entered Buckingham Palace .
and anything else he needed .	For the first time	in all his journeys he found a room that was
And that was a surprise .	For the first time	in her life she 'd allowed someone to hook on
needles through a frozen limb .	For the first time	in her life , perhaps , the Astropath had been
I was in great pain , and suddenly	for the first time	in my life , I forgot my fear of John Reed
actually looked at a flower , and	for the first time	in ten years he realized how beautiful
demented birds , disturbed	for the first time	in who knows how long , were battering
to hear that Matthew was dead .	For the first time	in his life , Matthew Cuthbert was an
her business on a sound footing	for the first time	in her life . ` It 'd be a grand thing
the elements as slaves , then	for the first time	in history slavery will be abolished . Human
a tankard . Blake was content	for the first time	in days . The barman returned , smiled , and
old man , propped in his chair	for the first time	in over a month , laid a trembling hand on the

Figure 3: The temporal pattern "for the first time in 'a long period of time'" (taken from Malá et al., 2021)

Analogously in Czech, 3/4-grams closely associated with the lemma *DOBA* (“time, period “) were found to comprise either the demonstrative pronoun *ta* (*té*, *tu*, ex. 2a, “that “) or the indefinite pronoun *nějaká* (*nějaké*, *nějakou,* ex. 2b, “some”).[71] The use of the anaphoric demonstrative is marked in Czech, as Czech does not have articles and normally does not explicitly mark the definiteness of nouns. Moreover, a comparison with the remainder of the general fiction corpus reveals that the colligation pattern *ta* + *doba* is slightly more frequent in the children's fiction subcorpus than in adults´.[72] This may suggest the importance of explicit anaphoric ties in the structure of children's narrative fiction (ex. 3) as well as the summarising effect of the temporal pattern with the demonstrative (ex. 4).

[71] Overlapping n-grams are merged in exx 2 a, b.

[72] 38% of noun phrases headed by *DOBA* include the demonstrative *TA* in children's fiction, while in general fiction (*SYN-7*) the proportion is 31%.

(2) a. (a) od té doby (se), od té doby, (co/než), (po) celou tu dobu, (./ale) do té doby

"(and) since that time (*se*-reflexive), since the time (when), (for) all that time", "(./but) till that time"

b. (.) po nějaké době, za nějakou dobu

"(.) after some time "

(3) Začnete za hodinu! Do té doby ať jsou připraveni tři nejdivočejší sloni.

"You start in an hour! By that time you are to prepare the three most fierce elephants. "

(4) Ve škole se teď nemluvilo o ničem jiném než o prázdninách. Zámožnější chlapci udivovali ostatní tvrzením, že pojedou do ciziny, a ukazovali ta místa na mapách. Jiní se spokojili s vesnicí nedaleko svého města . Byli také hoši, kteří neměli kam jet a byli odsouzeni strávit celé prázdniny doma. V této době příprav a neklidu položil Rikitan hochům [...] tuto otázku: "Nu, chlapci, co my budeme dělat o prázdninách?"

"Everyone at school spoke about nothing but the holidays. The more well-to-do boys amazed the others claiming they would go abroad, and pointed to those places on maps. Others made do with a village near their hometown. There were also boys who had nowhere to go and were destined to spend their whole holiday at home. At this time of preparations and unrest, Rikitan asked the boys [...] this question: 'Well, lads, what shall *we* do on holiday?'"

The patterns where DOBA is preceded by an indefinite pronoun, on the other hand, refer to periods of time of unspecified, yet not short, duration (ex. 5).

(5) Uplynula nějaká doba a kraj se změnil k nepoznání.

"Some time had passed and the area changed beyond all recognition."

Lastly, the temporal n-grams were analysed qualitatively, examining their contexts to determine their textual functions. Then I compared the patterns containing equivalent lexical nodes between the English and Czech data, focussing on the impact of typological differences. The findings of those analyses are discussed in the following section.

5.7 Analysis

5.7.1 N-gram nodes: nouns, adverbs, conjunctions

As mentioned earlier, the first step was searching through 500 top frequent lemmata in the respective Czech and English corpora to identify temporal expressions to be used as n-gram nodes. The selected nodes included nouns, adverbs and conjunctions. All nodes were employed as the basis for 3-4-grams, which in turn served as the starting point towards identifying temporal patterns, paying attention to the n-grams in context.

An examination of the resulting n-grams reveals that the part-of-speech of the node lemma has a major impact on the nature of the resulting pattern. There is a conspicuous difference between patterns with noun nodes on the one hand, and those with adverb or conjunction nodes on the other, in terms of their textual functions. The noun-node patterns typically comprise more than just one temporal expression, and consequently the temporal meaning extends over the entire pattern. For instance, patterns with *DAY* included *day and night; hours a day; one day when; day after day, every day, three days later*. In most n-grams with adverb or conjunction nodes, on the other hand, the temporal meaning stemmed merely from the node adverb or conjunction (e.g. *I´ve always wanted; when she arrived*); the particular time-denoting function is not associated with the whole sequence. Therefore, the following survey of textual functions is focussed on patterns with a nominal node. I compare the patterns containing equivalent lexical nodes between the English and Czech data, focussing on the impact of typological differences. (A quantitative summary of the textual functions of these patterns will be presented later in table 30 and discussed in detail.)

To ensure cross-linguistic comparability, I restricted the comparison to temporal nouns whose equivalents were represented in both languages. The resulting set of nouns (8 per each language) is presented in table 28. (Some nouns have more than 1 translation equivalent. Subsequently, I explored the functions of adverb and conjunction patterns. This complementary analysis will be presented later to offer a more comprehensive view of the scope of expressing temporal meanings.

English nodes	Czech equivalent nodes
DAY	DEN
MOMENT / WHILE	CHVÍLE
YEAR	ROK
NIGHT	NOC / VEČER
TIME	ČAS / DOBA
MORNING	RÁNO
EVENING	VEČER

Table 28. Correspondence of temporal nouns in the dataset

For practical reasons (to ensure manageability) I established an arbitrary cut-off point at the risk of n-gram value (i.e. collocation strength metric) of 50. The frequencies of patterns including the noun lemmata which have direct translation counterparts in the dataset are shown in table 29, in descending order.

ENGLISH	TIME	DAY	YEAR	NIGHT	MOMENT	MORNING	WHILE	EVENING	TOTAL
TOTAL	232	170	144	134	105	56	55	45	941
RISK ≥ 50	104	69	88	61	52	30	20	20	444
CZECH	CHVÍLE	DEN	ROK	ČAS	DOBA	NOC	RÁNO	VEČER	TOTAL
	“MOMENT“/ “WHILE“	“DAY“	“YEAR“	“TIME“	“TIME”	“NIGHT“	“MORNING“	“EVENING“	
TOTAL	256	242	143	91	76	74	50	36	968

Table 29. Pattern type frequencies for English and Czech

5.7.2 Textual functions of temporal patterns

In this subsection I outline the functional classification of temporal patterns, which applies both to the Czech and English data, i.e. patterns fulfilling these functions were attested in both languages. I have not identified any functions exclusive to one language only; however, individual functions may manifest themselves in different ways in each language. The lack of functions clearly limited to one language suggests that these functions are shared by texts intended for children regardless of language, being determined by the register characteristics. However, research into comparable corpora of adult fiction would be needed to verify this. The cross-linguistic differences seem to be grounded typologically and potentially also culturally; they will be discussed in the following subsection. I assigned only one (dominant) textual function to each pattern, while aware that some patterns may perform various functions depending on their context; or on the other hand that "a single occurrence" of a pattern "can be considered multifunctional" (Biber, 2006, p. 139). The patterns fulfil the following major narrative-organising functions:

- explicit reference to time (*at that moment*)
- emphasis, intensification (*day after day*)
- marking a turning point or a moment of suspense (*just in time to*)
- marking vague time (*once upon a time*).

Another factor that was observed, overlapping with the textual functions, is idiomaticity. This was understood in broad terms, to encompass invariable non-compositional patterns which carry non-literal meaning.

Table 30 shows the numbers of pattern types fulfilling each function (within the limited sample of patterns with risk of n-grams value ≥ 50), as well as the number of idiomatic patterns. Some patterns were not found to fulfil any of these functions – their numbers are listed under "other".

ENGLISH:							
function	explicit time	intensification	suspense	vague time	other	TOTAL	idiomaticity
TIME	44	31	4	7	18	69	4
MOMENT	4	7	38	2	1	52	2
WHILE	0	2	12	1	5	20	1
DAY	32	8	0	26	3	88	4
YEAR	67	10	0	0	11	61	3
NIGHT	39	8	0	1	13	104	4
MORNING	23	1	1	1	4	30	1
EVENING	11	1	0	3	5	20	1
TOTAL	220	68	55	41	60	444	20
CZECH:							
function	explicit time	intensification	vague time	suspense	other	TOTAL	idiomaticity
ČAS "TIME"	1	9	13	8	14	56	8
DOBA "TIME"	27	15	0	0	18	49	1
CHVÍLE "WHILE"/ "MOMENT"	25	20	47	12	0	104	3
DEN "DAY"	62	20	18	0	11	32	8
ROK "YEAR"	59	13	7	0	6	26	2
NOC "NIGHT"	9	9	3	0	4	83	4
RÁNO "MORNING"	13	7	0	0	7	107	4
VEČER "EVENING"	3	5	0	0	6	14	3
TOTAL	199	98	88	20	66	471	33

Table 30. Textual functions of patterns[73]

[73] By n-gram-type frequency, risk of n-grams ≥ 50. Idiomaticity is an additional feature,

The individual functions will now be discussed and illustrated by examples, ordered by their frequency in the English data. To achieve a comprehensive portrayal of the functions, the examples also include selected temporal patterns with nodes which were identified in the data but not included in the crosslinguistic comparison, hence are not included in table 30.

5.7.2.1 Explicit reference to time

Various aspects of temporal relations are made explicit by the patterns:

a) Specific temporal reference to a moment or punctual action, as in ex. 6. This action may be repeated, as in *three times a day* – ex. 7.

(6) Mojsus si krabičku vzal, v tu chvíli však žádný z nás netušil, že v ní není mince od Tibeťanky, ale úplně jiná.
"Mojsus took the box, but at that moment none of us knew that in the box there was not the Tibetan´s coin but an entirely different one."

(7) The room in the workhouse where the boys were fed was a large stone hall, and at one end the master and two women served the food. This consisted of a bowl of thin soup three times a day, with a piece of bread on Sundays. The boys ate everything and were always hungry.

b) Alternatively, the pattern serves to explicitate the duration of an action, contributing to structuring the narrative and/or building suspense. In example 8, the expression *po nějaké době* ("after some

occurring across the textual functions.

time") explicitly marks the (limited) duration of the process.

Other examples of this function include the patterns:

- for a moment, for a little while, for a long time;
- po nějakém čase, hodnou/hezkou chvíli, čas plynul ("after some time", "a nice/pretty time", "time passed").[74]

(8) Smutné tři dcerky se přesto staraly o ubohého ptáka, pravidelně mu dávaly jídlo a pití, a po nějaké době kačerovi znovu narostlo peří.
"The three sad daughters cared for the poor bird anyway, regularly giving it food and drink, and after some time the gander grew feathers again."

c) Other examples contain explicit marking of temporal relations between two or more actions, happening either simultaneously or in sequence. Examples include:

the next morning, but after a while, for the first time, at that moment; at the same time, after a moment's hesitation;

po nějakém čase ("after some time"), nastal čas ("the time has come"), tak – a teď ("right – and now"), když nadešel čas ("when the time came"), jednoho dne ("one day"), teprve teď ("only now"), hned zítra ("right tomorrow").

(9) The doctor decided to operate on my eyes, and the next morning we got up early.

(10) Lavril potom vyndala ze své brašny obvazy i zklidňující roztok. „Tak, a teď mu budete muset ten šíp vytáhnout," pokračoval kentaur.
"Lavril then took some bandages and a calming tonic out of her

[74] The same examples are also cited in Malá et al. (2021).

satchel. 'Right, and now you will have to pull that arrow out for him,' the centaur went on."

5.7.2.2 Intensification

The patterns associated with intensification refer to the passing of time or to repeated happenings. In both languages, intensifying patterns contain parallel adverbial structures consisting of noun or adverbial phrases linked by coordinators or prepositions. The phrases are often repeated, as in ex. 11 and the following examples:

day after day, day by day, once and for all, day and night, over and over again;

ve dne v noci ("day and night"), večer co večer ("evening by evening"), čas od času ("from time to time"), den ze dne ("day by day").[75]

(11) Víme jen tolik, že princezna večer co večer vysedává u otevřeného okna a čte si v zelené knížečce, kterou jí přivezl sám otec král.
"All we know is that evening after evening, the princess sits by her open window and reads from a green booklet that her father King gave her."

5.7.2.3 Marking a turning point or building suspense

As observed by Thompson and Sealey (2007), multi-word patterns are used in children's fiction to mark a suspenseful moment or a turning point in the narrative. These patterns typically contain or co-occur with intensifiers or focalisers, such as *just*, *only*, *právě* ("just"):

just in time to, then/when suddenly, at this very moment, at that moment (there appeared), too late;

[75] (These examples are also cited in Malá et al., 2021).

právě v tu chvíli ("just at that moment"), právě ve chvíli (, kdy) ("just at the moment (when)"), nemáme moc času ("we don´t have much time"), (ne)ztrácet čas ("(not) to lose time"), každou chvíli ("every now and then" / "any minute now"), je nejvyšší čas ("it is high time"), v poslední chvíli ("at the last moment").

Examples 12 and 13 provide wider contexts which illustrate the focalising function these patterns fulfil in the text.

(12) Mary gave one horrified glance at it, then flung herself madly into the ditch at the side of the lane. She was only just in time to escape being knocked down.

(13) Rytíř stačil jen v poslední chvíli uskočit a odrazit vidličku mečem. "At the last moment the knight <u>just</u> managed to jump aside and fend off the fork with his sword."

5.7.2.4 Vague temporal reference

A number of patterns express vague temporal meanings, typically referring to the past. The indefiniteness of temporal reference contributes to the construal of the narrative as unreal or supernatural. This is in line with Knowles and Mjalmkjaer´s observation that "fairytale writers may facilitate our perception of our own world as magical by refraining from explicit spatiotemporal staging, or by providing an impression of temporal and spatial distance between reader (and writer) and story." (1996, p. 160)

Examples of vague time reference include *once upon a time, in those days, a long time ago; one evening, many years ago; před dávnými časy/věky/léty/lety* ("many ages/years ago"), *za dávných časů/dob* ("in times long past"), *jednoho dne*, ("one day"), cf. ex. 14. Some of these patterns are markedly register-specific (*once upon a time*[76]), or their

[76] Note: *once upon a time* occurred outside our analysed sample (risk of n-gram = 40).

usage may be evocative of fairy stories (*před dávnými léty* – "long time ago").

(14) Před dávnými časy žil v jedné vesnici v Kašmíru mudrc,…
"Long time ago in a village in Kashmir there lived a wise man…"

5.7.2.5 Idiomatic expressions

Idiomaticity was found to occur across the individual functional pattern groups. Idiomatic patterns are not restricted to a single function; rather, they seem to play the role of style markers, associated with the aesthetic function of the text. Likewise, they may contribute to the educational function of children's fiction (cf. Čermáková & Chlumská, 2016), providing the young readers with varied lexical input so as to expand their vocabulary knowledge.

Idiomatic patterns formed a highly heterogeneous group, represented mostly by patterns with adverbial functions, such as *day and night*, *spur of the moment*, *from time to time*, *once in a while*, *hang on a minute*; *den ze dne +* comparative ("day by day + comparative"), *hodnou/hezkou chvíli* ("a pretty long time", lit. "a nice moment"), *do roka a do dne* ("in a year and a day's time"), *dočkej času jako husa klasu* (proverb roughly equivalent to "Rome wasn't built in a day", lit. "wait for your time like a goose does for her ear of wheat"), *do nejdelší smrti* ("forever and a day", lit. "until the longest of deaths"), *od rána do večera* ("from morning to evening").[77]

5.7.3 A contrastive look at temporal patterns

The differences between English and Czech patterns seem to be grounded in the typological characteristics of the languages. The most salient differences are related to non-correspondences in n-gram length.

[77] All translations in this paragraph are mine.

English employs articles, which causes noun-based n-grams to be longer than their Czech equivalents (*za chvíli – after a while*). Other length non-correspondences are caused by the fact that Czech, being synthetic, prefers to express grammatical meanings as well as some lexical modifications by suffixes, e.g. an English preposition occurs in lieu of a Czech case ending (*chvíli/chvilku* “moment-accusative” *– for a moment, for a (little) while*). Similarly, the diminutive meaning is expressed by a suffix in Czech, corresponding to an English adjective (*za chviličku* “after while-diminutive” – *a little while later*).

Some differences can also be observed in the composition of n-grams. N-grams of a comparable length may comprise a personal pronoun in English, while their Czech equivalents contain a verb form, or the pronoun *se*, a constituent of reflexive verbs in Czech. The Czech verb expresses the grammatical categories of number and person in its personal ending: *za/po chvíli byli/bylo/se/řekl* “after while they-were/it-was/*se*-reflexive/he-said”*.* In English, the categorial meanings of number and person are expressed separately through personal pronouns, *after a while (he/she/I/the)*, which are typically followed by a lexical verb, as in *after a while he stopped/said/began.* Thus, we are ultimately dealing with another case of n-gram length non-correspondence.

Some Czech patterns contained explicit anaphoric or cataphoric links, as in *od té doby, co* (*since the time when*). While these means of expression are marked in Czech and enhance the cohesion of the narrative, in English such links in the form of determiners are omnipresent given its analytic nature.

Another typologically based difference consists in word order within the n-grams. Apart from the positional variability in Czech, some patterns illustrate the different positions of modifiers. In *právě/zrovna v té/tu chvíli* “exactly/just at that moment”, the focussing adverbs *právě* or *zrovna* premodify the whole adverb phrase *v té/tu chvíli.* These patterns correspond to the English *at this/that very moment*, where the intensifier *very* premodifies only the head noun. Different dependency relations in the noun phrase may relate to word order differences, as in the pattern *po chvíli ticha* ("after a moment silence-genitive") corresponding to *a moment's silence*.

To sum up, the similarities between the patterns in both languages appear to stem from their textual functions, associated with the register of children's narrative fiction. The differences can be ascribed to linguistic differences between English and Czech, and – in the case of idiomaticity – perhaps to the culture-specific features of the register.

5.7.4 Idiomatic patterns

As noted previously, (broadly conceived) idiomaticity seems to play an important role in the register of children's literature. In English, idiomatic patterns comprised 4% of the patterns analysed (20 out of 444 patterns with noun nodes), and in Czech 7% of the sample (33 out of 471 patterns; cut-off point for both languages = risk of n-grams ≥ 50). The 3% difference between English and Czech is statistically significant at p = 0.0476. This suggests that our Czech texts employ a larger number of idiomatic expressions than English overall.

In Czech, some of the idiomatic pattern types overlap, forming longer patterns together (*dočkat času jako + času jako husa = dočkat času jako husa klasu*). Other types can be considered variations on one pattern, such as *právě/zrovna v tu/tuto chvíli* - "just at that/this moment"; or *být/mít nejvyšší čas, aby* - "be/have high time to". By contrast, for English idiomatic patterns, no such overlaps between pattern types were identified. Apparently, the idiomatic patterns attested in my Czech corpus are all – though to varying degrees – typical of fiction intended for children and teenagers. This is indicated by comparing their respective relative frequencies in children´s literature as opposed to adults'. Their i.p.m. rates are consistently higher in the children´s fiction subcorpus.

Likewise, the English data contained idiomatic multi-word units. At first sight most of them may seem stylistically neutral and not typical of a particular register (there are no expressions stereotypically associated with fairy-tales such as *once upon a time*, and fewer instances of "colourful" idiosyncratic language overall than in Czech). However, a look at the textual metadata reveals that a number of them are also more frequent in fiction for young readers than in adults'. Nevertheless, the results are less conclusive than those for Czech data. This is partly due to some sections of the BNC lacking the metatextual information on target audience. Moreover, although the i.p.m. of some examples, such as *spur of the moment*, suggests that the expression may be more typical of children's fiction, the raw frequencies are very low, making this conclusion rather tentative.

In summary, the idiomatic temporal noun patterns identified in our data were found to be more frequent in children's books than in fiction written for adult audience (table 31); the absolute frequencies, however, are too low to be conclusive, and the quantitative results should be accompanied by a qualitative functional analysis of the phrases in adult's fiction before any conclusions can be drawn. The frequencies listed in table 31 are of the 5/6 most frequent idiomatic patterns in children's (corpora *SYN-7-jun*, *BNC-jun*, see table 26 above) and adults' fiction subcorpora of *SYN-7* and *BNC*.

English (BNC)	children's fiction		adults' fiction	
	total hits	ipm	total hits	ipm
just in time	22	10.749	144	8.315
time to time	19	9.283	308	17.785
day and night	15	7.329	60	3.465
in no time	10	4.886	93	5.370
all day long	9	4.397	34	1.963
day after day	8	3.909	59	3.407

Czech (SYN-7)	children's fiction		adults' fiction	
	total hits	ipm	total hits	ipm
od ráno do večera	33	11.7	231	6.99
čas od času	31	10.99	580	17.55
nejvyšší čas	30	10.63	293	8.87
dnem i nocí	14	4.96	58	1.76
právě v tu chvíli	8	2.84	32	0.97

Table 31. The frequency of idiomatic temporal noun patterns in children's and adults' fiction (ipm = instances per million tokens)[78]

5.7.5 Conclusions

My methodological research question concerned the potential assets and pitfalls of employing n-gram extraction in contrastive analysis. Addressing issues encountered in the two n-gram-based case studies presented in the previous chapters, I have amended the parameters of the n-gram search employed in the present research as follows: N-grams were extracted from lemmatised data to accommodate morphological variability; n-grams composed of the same lemmata differing only in order were clustered and an open slot was enabled to allow for positional variability; shorter n-grams which were contained in a longer gram were grouped with it to reflect the fact that "shorter bundles are often incorporated into more than one longer lexical bundle" (Biber et al., 1999, p. 990).

First a frequency list was used to identify frequent temporal expressions, subsequently these were used as the basis for n-gram extraction and I proceeded towards temporal patterns with the help of

[78] The sizes of the subcorpora of original fiction written for adult target audience from SYN-7 and BNC are 33,047,774 tokens / 780 texts, and 17,317,696 tokens / 376 texts, respectively).

qualitative analysis of the n-grams in context. The *Engrammer* software, allowing for lemmatisation, collapsing n-grams of varying lengths, and the extraction of unordered n-grams, helped overcome some of the caveats stemming from the typological differences between English and Czech (i.e. non-correspondences in n-gram length and word order). The present study has confirmed that n-grams can be used as a convenient stepping stone towards the identification of patterns, allowing for automatic extraction of a large amount of data. Still, the application of n-grams to inflectional languages like Czech has its limitations. A contrastive approach offers a suitable complementation to the n-gram method, revealing the ways in which n-grams are influenced by typological characteristics of the respective languages.

Within the node lemmata which were used as the basis for the identification of patterns, there proved to be a major difference between individual word classes. In patterns based on temporal adverbs or conjunctions, the temporal meaning tends to be limited to the node adverb/conjunction only, e.g. *'ve <u>always</u> wanted, <u>always</u> too afraid, <u>when</u> he saw me*. Only few typical temporal patterns with specific textual functions based around an adverb or conjunction node have been identified (e.g. *then/when suddenly I*). On the other hand, nouns occur in patterns with specific temporal functions which are often not derivable from the noun node itself. Further, using the *Engrammer* software has enabled me to identify variable temporal patterns in which different temporal nouns may alternate in a given slot, e.g. *for the first time/couple of weeks/six months; day/morning and night; hours a day/month/week; in those days/months/weeks.*

The patterns obtained through the n-gram-based analysis of children's and teenagers' fiction fall into several functional groups. While referential patterns were prevalent, the corpora also included patterns serving text-structuring functions, such as signposting relevant events in the narrative and relating individual happenings and actions to each other.

A closer examination of the patterns in context then helped me identify their particular functions in the narrative texts. Their major functions were explicit signalling of time, intensification, marking a turning point in the story, and expressing vague time (including highly register-specific items such as *once upon a time*, serving as instantly recognisable and conventionalised register markers).

The results suggest that time in children's fiction is indeed organised in specific ways. Further, the qualitative examination of the patterns identified has indicated that children's fiction is marked by a considerable degree of idiomaticity, which had not been expected: some of the idiomatic patterns are register-specific (e.g. *once upon a time*). This may suggest that idiomatic language is used purposely with regard to the intended audience and their specific needs. Since children's literature fulfils an aesthetic as well as educational role (Čermáková & Chlumská, 2016, p. 163), the authors may be employing rich idiomatic language in order to attract the readers' attention as well as to help develop their vocabulary.

A possible caveat lies in the influence of idiosyncratic authorial idiolects. As Gray and Biber have pointed out (2015, p. 137), there is a substantial "influence of corpus design on the identification of important lexical phrases". For example, the Czech children´s fiction subcorpus contained 4 instances of the idiomatic pattern *to BÝT doba, než* ("it TAKE a long time before", lit. "that BE a time before"), all from the same source text *Školák Kája Mařík* by Felix Háj. Another case in point may be represented by examples from texts by authors with a highly idiosyncratic style, e.g. Karel Čapek, whose work is part of the literary canon included in school curricula and therefore easily recognisable for most Czech readers. In order to compensate for this potential idiolect effect, a larger and more diverse corpus would be needed.

Finally, the scope of the research could be expanded by conducting a similar study on parallel rather than comparable corpora, examining

Czech-English translation equivalents of temporal n-grams. Another worthwhile research question would be to what extent the patterns identified in the children´s fiction corpora also occur in adults' fiction and whether they perform similar functions there. As discussed earlier, a preliminary small-scale probe into the occurrence of idiomatic patterns in children´s as opposed to adults´fiction has suggested that the two registers may display interesting differences in this respect.

6 Conclusions

This corpus-driven study has addressed phraseology from a register- and cross-linguistic perspective. The first chapter introduced the frequency-based approach to phraseology, sketching its development and major trends within the field. The second chapter discussed methods in frequency-based phraseology and pointed out potential problematic areas associated with them. The following three chapters presented the results of three n-gram-based case studies, each focussing on a different register, offering a glimpse into the phraseology of parliamentary debates, newspaper reporting and children´s fiction, respectively. In each case study the focus was on a selected area of phraseology, using different variations on the n-gram method: gradually adjusting the parameters of n-gram size, lemmatisation, and variability within n-grams (order and empty slots). With every study the method was adapted in an attempt to address methodological issues encountered in the previous studies. To learn more about the functions of patterns in context, n-grams were complemented by collocations (case study II) and with lexemes from a particular semantic field (case study III). Furthermore, the register phraseologies were consistently compared between English and Czech, aiming to identify how phraseological differences are shaped by the typological features of the two languages.

The present chapter summarises the findings and attempts to draw broader conclusions in all three respects observed throughout the three case studies. Firstly, I will outline the phraseological characteristics of the three registers which were identified with the help of n-grams. Secondly, I will comment on the cross-linguistic comparison of English and Czech phraseology which was obtained through the lens of the three register phraseologies. Thirdly, I will evaluate the n-gram method and its variations which were employed in the case studies, discussing problematic areas and considering the applicability of the method to English and Czech. Finally, several areas warranting further research will be suggested.

6.1 Register phraseologies

This section summarises the findings from the three case studies and subsequently attempts to generalise about what features of register phraseology may be revealed through n-grams.

6.1.1 Parliamentary debates: an experimental approach to n-gram length

The first case study was purely corpus-driven: it experimented with n-gram length, comparing what features of the parliamentary register were revealed by different lengths of n-grams, and attempting to determine which lengths were the most suitable, i.e. which n-grams would yield the most informative results. The parliamentary debates are a highly specialised register; hence we may expect recurrent patterns fulfilling communicative functions frequently required by this register. N-grams are therefore expected to be a suitable method to reveal such functional patterns.

The comparison of n-gram lengths between 2 and 10 revealed that each length pointed to different discourse functions. For instance, short n-grams highlighted the importance of deixis in the debates – interpreted as reflecting the need for precision and clarity; while long sequences pointed to fixed formulae or whole sequences of patterns, some of which were highly register-specific. The occurrence of such long stable sequences was seen as a consequence of the register being highly specialised and governed by strict norms, to which the speakers conform in an effort to maintain their credibility and legitimacy within the community.

The initial aim to determine a specific n-gram length as the optimum one was found to be problematic: each length of n-grams may be valuable in revealing a particular aspect of the recurrent patterns, and combining several lengths consequently allows for a comprehensive view of the register´s phraseology. This suggests that a combination of several different n-gram lengths is the most fruitful, as it reveals a wide range of patterns with various functions, but the downside of this approach is the impossibility of reaching precise quantitative data. This is due to numerous overlaps between n-grams of different lengths: many shorter n-grams are contained in longer ones, i.e. some sequences occur repeatedly and as a result are overrepresented. The precise extent of this repetition or overrepresentation cannot be determined. Hence, extracting n-grams of several lengths separately results in a broad view of the phraseological patterns in the data, the results being limited to purely qualitative observations. The study also points to the need for overlapping n-grams of different lengths to be collapsed (addressed in case study III with the help of Engrammer, Milička, 2019).

An interesting finding, which however falls beyond the scope of this thesis, was related to the Speaker´s turns in the parliamentary debates: they were mostly limited to one-word utterances, as in the case of *order*. This single word proved to play an important and register-specific role in the data: it structures the debates and is associated with the speaker's being in charge of the discourse. Obviously, n-grams fail to identify such single-word turns. This points to the importance of looking beyond phraseological means, if a register is to be characterised in a comprehensive way.

Apart from patterns with other functions (such as honorifics and address terms, content patterns referring to political entities, or modal patterns, to name only a few), a number of patterns identified in the parliamentary debates served a text- or discourse-structuring role, be it deictic patterns for intratextual reference, or performative formulae structuring the parliamentary session. This prompted a closer investigation of structuring patterns in the second case study (prepositional patterns).

The first case study was intended as an experimental probe, examining a highly specialised register, and using a small dataset – the results suggested that even a small amount of data can be sufficient for revealing recurrent patterning (in line with the findings of Gries et al., 2011). By contrast, the following two studies use larger corpora with representativeness in mind, and the registers there are not as narrowly functionally defined as parliamentary debates: newspaper reporting and children´s fiction. Since these registers are less specific or *niche*, the n-grams in them tend to be less repetitive than those found in the parliamentary debates.

6.1.2 Newspapers: prepositional patterns and evaluative meanings

The second case study is devoted to the phraseology of newspapers. As an initial step, a pilot study was conducted, restricted to British newspapers. However, the exploratory study design adopted in this pilot was not found to yield satisfactory results: the findings were heavily affected by the presence of topics which were overrepresented in the press, especially in tabloid newspapers (patterns related to sports reports). Still, the results of the pilot study hinted at some interesting areas, the potential (under)representation of female speakers being a particularly inspiring suggestion for further inquiry.

As the exploratory pilot study proved problematic, for the contrastive case study proper, a "lexis to n-grams" (Stubbs 2007b, cited in Lindquist & Levin, 2008, p. 145) approach was adopted instead. As Groom (2010) argues in favour of grammatical words to be used as a springboard towards identifying frequent and pervasive phraseological elements, I chose to use selected frequent prepositions as n-gram nodes, aiming to identify patterns with text-structuring functions, which were identified earlier in case study I.

A new aspect which was introduced in case study II is lemmatisation. Using lemmatised data is envisaged to help bring together n-grams differing only in inflectional affixes. Further, n-grams are complemented with collocations: by examining the collocations of frequent patterns, we may reveal how those patterns operate in context, how they contribute to building meanings throughout texts. This study focussed particularly on the role of prepositional patterns in expressing evaluative meanings.

The prepositional patterns identified in English and Czech newspapers reflected both the primary informational dimension of the newspaper texts (referring to past and present events, quantification, quotations), as well as having additional evaluative meanings: some patterns were found to have an opaque evaluative prosody, or a particular semantic preference.

6.1.3 Children´s literature: temporal patterns

In case study III the focus is on children´s fiction, a register characterised by a complex functional load – serving entertainment as well as educational purposes, contributing to the young readers´ linguistic but also social development. Similarly to case study II, a "lexis to n-grams" approach (Stubbs 2007b, cited in Lindquist & Levin, 2008, p. 145) is employed. In case study II, this involved a selected word class, namely prepositions – function-word patterns were chosen as they are able to reveal pervasive phraseological characteristics of the register at hand (Groom, 2010). In case study III, to complement the previous approach, the n-grams extracted are based around a different kind of node lemma: here the node is defined in semantic terms as a temporal expression. The nodes were identified with the help of a frequency list: among the most frequent lemmata in the corpus, temporal ones were selected and employed as n-gram nodes. A further restriction was then imposed, focussing on patterns found around temporal nouns: this was based on the finding that in nominal patterns, the temporal meaning was expressed by the whole phraseological sequence, while in patterns containing a temporal adverb or conjunction the temporal meaning was condensed in the node itself. This suggests that on a "lexis to n-grams" approach, the word-class of the lexical n-gram node should be carefully considered.

The method was further adjusted by using lemmatised data, and above all by employing the Engrammer software (Milička, 2019) to introduce positional variability to n-grams. Specifically, the software allowed for an optional slot within n-grams, and collapsed n-grams which differed only in word order. This n-gram search revealed that temporal patterns fulfilled four major functions in children´s fiction: explicit temporal reference, intensification, marking a dramatic moment, and vague time reference (highly register-specific patterns such as *once upon a time* were found in this category). Idiomaticity was also observed in a number of the patterns – further research focussing on idiomatic patterns could point to interesting findings about the language of children´s fiction. The use of idiomatic patterns is possibly motivated by an effort to enrich the young readers´ vocabulary. Comparing the occurrence of idiomatic patterns between fiction for children and adults would reveal whether idiomaticity is consistently more prevalent in children´s books, as was indeed suggested by the results of study III. The findings further suggested that idiomaticity was more frequent in the Czech data, which would likewise warrant further investigation.

Patterns with text-organising functions (which had largely been the focus of the previous case study II) were also represented in children´s literature, expressing explicit temporal relationships between actions and structuring the narrative. Overall the results from all three registers pointed towards such patterns, suggesting that discourse-organising functions tend to be fulfilled by conventionalised phraseological means.

Finally, a potential caveat was seen in the effect of authorial idiolects, as recurrent n-grams may reflect expressions characteristic of a particular writer, whose language would then be overrepresented in the patterns identified. In the future this can be resolved by carefully balancing the set of authors whose works are represented in a corpus.

6.2 Cross-linguistic findings: Czech and English

Throughout the three case studies, n-grams have reflected the typological characteristics of Czech and English. Some of these characteristics make cross-linguistic comparison problematic, as indicated by the first case study, a corpus-driven probe using various n-gram sizes. This is especially evident in short n-grams: while English bigrams comprise mostly function words, reflecting its analytic nature, Czech bigrams contain lexical words. The comparability of different n-gram sizes across the two languages is therefore disputable.

In the newspaper register, prepositional patterns containing the translation equivalents *v* and *in* allowed for a cross-linguistic comparison. Patterns with both prepositions occurred with a similar frequency (1,392,147 *in* patterns, 1,179,492 *v* patterns), *v* patterns being more varied (167 *v* n-gram types as opposed to 75 *in* ones). Prepositional patterns which were amply represented in newspapers in both languages included adverbial patterns of time or place, patterns carrying quantitative meanings or referring to particular points in time (*in the first, second, last*). Some differences were observed between Czech and English temporal patterns: Czech newspapers contained more references to planned events (*dnes v ___ hodin*), while in English we saw more of temporal patterns referring to broadly conceived spans of time (*in the recent years*). Czech was found to employ more varied verbal patterns, many referring to past and future events (*se uskutečnit v* "take place in"). The findings indicated that Czech newspapers use the examined prepositional patterns to introduce sports or cultural events. While such uses were not observed in the English data, this may be simply because English uses different prepositions to convey those meanings, as uses of preposition translation equivalents often vary between languages (Klégr & Malá, 2009).

In children´s literature, the main difference between English and Czech was found to lie in a non-correspondence of n-gram lengths: in case of pattern pairs expressing equivalent temporal meanings, the English n-grams tended to be longer due to the use of function words (*za chvíli – after a while*), while Czech employs affixes to convey grammatical meanings (case endings corresponding to English prepositions; verb suffixes) as well as lexical meanings, specifically through diminutive suffixes, which can be expected to occur in children´s fiction due to being emotionally expressive (*za chviličku*). Overall, the Czech and English patterns in children´s fiction were found to fulfil similar textual functions, as required by the register at hand; while differences observed between the patterns were due to the typological nature of the languages.

Finally, cross-linguistic comparison may point to differences grounded not just in language but also in its cultural background. In the parliamentary debates, some patterns clearly reflected local cultural norms, especially related to politeness strategies: this was shown by the honorific terms of address employed in the British parliament, while no equivalent address terms were attested in the Czech proceedings. More research would be needed to specify whether some of the differences between Czech and English patterns in the other two registers may also reflect any cultural specifics. Potential areas for further inquiry in this vein include the frequency of reference to sports/cultural events in newspapers, and the prevalence of idiomaticity in children´s fiction. Another interesting area for cross-linguistic and cross-cultural comparison would be the linguistic representation of speakers of different genders, especially in newspaper reporting. As was identified by function word patterns in the pilot study of English newspapers, female speakers were less frequently quoted in the newspapers than male speakers. In the Czech newspaper corpus, the collocations of the complex preposition *v čele s* indicated that women performing a leading role (e.g. *předsedkyně, brankářka*) are referred to less often than their male counterparts.

6.3 Methodological findings

The n-gram method was employed in three case studies. In each study it was adjusted and combined with different approaches. This section summarises and evaluates the individual modifications which have been applied to n-grams in the previous chapters.

Lemmatisation was introduced as it meets the need (especially apparent in Czech) for collapsing n-grams which only differ in inflection. However, in case study II we saw that the utility of lemmatisation is debatable. First, some lemma-n-grams were found to be identical with word-form n-grams, as in multi-word prepositions. This suggests that when focussing on function-word patterns, lemmatisation may be less relevant than in the case of lexical word patterns. Second, lemmatisation may obscure differences between patterns. This was illustrated on the lemmatised pattern *v tento den*: While the more frequent plural form *v těchto dnech* (pl.) was used to refer to topical affairs in progress over the present period of time but also as a vague time marker referring to recent events, with the focus presumably on the event rather than on precise timing, *v tento den* (sg.) referred to a specific day. Hence, when extracting n-grams from lemmatised data, the particular word-form-n-grams should remain accessible – comparing the uses of different word-forms-n-grams may point to two different meanings, i.e. in our understanding to two different patterns.

Allowing for positional variability within n-grams was another parameter which was examined and found to be useful in collapsing overlapping n-grams, as well as shorter n-grams contained in longer ones, which had posed major problems in case study I. Specifically the positional variability was of two kinds: empty slots inside n-grams, and variable word order (shuffle-grams) – both were applied in case study III thanks to the Engrammer software (Milička, 2019).

In sum, introducing lemmatisation and positional variability had a similar effect: these two parameteres help collapse n-grams which differ only in a minor formal aspect (inflection or word order). The accessibility of a breakdown to the individual pre-collapse variants is key, as comparing the uses of different word-form n-grams may point to different patterns. This seems to be a viable approach, tapping into the benefits of lemmatisation whilst being able to balance its potentially undesirable effect of obscuring the differences between patterns.

Throughout the three case studies, the results have suggested that n-grams are useful in initial stages of research, as they enable us to scan a large amount of data for recurrent patterning; but they seem to work best when combined with other methodological steps in order to refine the findings, make them focussed and more informative. As shown in case study II, combining n-grams with collocation can help us describe more precisely how recurrent phraseological patterns (identified by n-grams) work in context and contribute to building large-scale textual meanings (revealed by collocations). Further, we may search for n-grams containing a particular word or lemma. This approach was illustrated using grammatical words (case study II) and subsequently with lexical words (case study III). N-grams containing grammatical words – namely prepositions – efficiently identify patterns with text-organising functions which are evenly dispersed through the corpus, which ensures that we obtain a comprehensive portrait of the phraseology of a given register. This aspect of even dispersion cannot be ensured by n-grams based around a pre-selected lexical word. However, these lexical-word patterns can capture phraseology referring to a particular semantic area, in this case the depiction of time. The study of lexical-word patterns has indicated that different word classes may be associated with different phraseological patterns. In patterns centred around nouns the temporal meaning was expressed holistically by the whole pattern, while in adverb patterns the temporal meaning stemmed from the adverb only, suggesting that possibly, the idiom principle is manifested differently in individual word classes. This issue would make an interesting topic for further inquiries.

Finally, two experimental steps were incorporated into the n-gram searches at different stages of the research. In the initial case study, extracting n-grams of various lengths indicated that different n-gram lengths point to different phraseological aspects of the register at hand. This may prove useful, especially for a highly specialised register, such as the parliamentary debate transcripts in case study I. However, the downside of this experimental approach is that the results are to a large extent repetitive – and the precise extent of this repetitiveness cannot be determined, as demonstrated in case study I. A related drawback of n-grams is that they seem prone to highlighting and even overrepresenting semantic areas where the language is highly repetitive (which may be for various reasons: either the language in a particular register is highly standardised, as in parliamentary debates; or the subject matter is more or less constant and does not allow for much variation, as in sports newspaper reporting, which was overrepresented through the n-grams).

Secondly, including punctuation in n-grams (as seen in case study II) was another experimental step which has proven interesting. In the study of newspaper register, punctuation pointed to a tendency for patterns to recur around syntactic boundaries. It also reflected cross-linguistic differences, as Czech subordinate clauses are preceded by an obligatory comma, which often has no counterpart in English. In future studies, when allowing for punctuation in n-grams, it may be useful to consider extending the n-gram length by one position: since the punctuation marks occupy a slot, we effectively retrieve a shorter sequence of words.

6.4 Suggestions for further research

The results of the present study have pointed towards several areas which warrant further research. Firstly, from the crosslinguistic perspective, it would be valuable to apply the n-gram methodology to more languages, especially inflection-rich ones, and compare them to the results from Czech. Adding further languages to the comparison would also serve to further test the effects of using lemmatised data for n-gram extraction.

Apart from the cross-linguistic outlook, this study examined phraseologies along the register dimension. Therefore, examining phraseologies in other registers would be a suitable complementation to the present study; especially functionally specialised registers distant from those analysed here, such as the language of social media, which seems to date rather underresearched from the phraseological perspective. Another possible extension would be to study phraseology in a spoken register, which could be compared to case study I.

As mentioned in case study III, comparing children´s fiction to fiction for adult readers would be a suitable next step to complement the present results. Comparing the phraselogical characteristics of both kinds of fiction using keyword analysis would allow us to map the extent to which the patterns identified in children´s fiction are typical of writing for children. Further, to complement the cross-linguistic outlook of the present case study II, it would be worthwhile to explore phraseology in children´s literature using a pair of parallel corpora, identifying Czech and English translation counterparts of frequent patterns.

Another area of research which was beyond the scope of this study but would undoubtedly deserve scholarly attention is the study of phraselogy in a broader cultural context, i.e. examining how phraseology reflects cultural phenomena. Children´s fiction would be one suitable register to this end, as it has an important educational and socialising function, introducing young readers to values and cultural norms. Arguably, other registers likely to reflect cultural differences also include newspapers; hence a cultural perspective on register phraseologies may be an interesting extension of the present study. Furthermore, a cross-linguistic study of phraseology would be especially promising, as it allows for identifying potential cultural differences. In a related vein, n-gram analysis may prove useful for the purposes of critical discourse analysis: n-grams alone highlight recurrent features of texts, making them promising starting points for CDA; and as shown in case study II, the collocations of frequent patterns can then show how those patterns are involved in construing meanings throughout texts through their semantic prosodies and preferences. N-gram-based studies can therefore contribute to the line of research represented by CADS (corpus-assisted discourse studies, Partington, 2010). In a related vein, a diachronic study of the development of semantic prosodies and preferences of patterns over time would be a worthwhile next step: it may shed light on how positive or negative evaluative meanings of frequent collocates gradually become ingrained in the node pattern, in a process which is possibly akin to Croft´s notion of hypoanalysis,[79] which relates to syntactic properties (2000, p. 127):

> In hypoanalysis, the listener reanalyzes a contextual semantic/functional property as an inherent property of the syntactic unit. In the reanalysis, the inherent property of the context [...] is then attributed to the syntactic unit, and so the syntactic unit in question gains a new meaning or function.

In case of semantic prosody there would be a transfer of an evaluative meaning from frequent collocates into the node pattern. This would suggest that the negative evaluative meaning of collocates can gradually become part of the pattern´s connotational meaning.

As regards potential practical applications of the n-gram method, n-gram-based analysis of phraseologies centred around grammatical words may provide valuable input for teaching practice, as suggested in Vašků et al. (2019). Phraseological competence has been shown to play an important role in a speaker´s proficiency (Hyland, 2008; Paquot, 2018; Paquot & Granger, 2012), and pose problems to L2 learners, including advanced speakers (Granger & Bestgen, 2014, p. 229). Learners tend to overuse a limited number of patterns (Granger, 2017; Hasselgård, 2019). This obstacle may be overcome by dedicating more classroom time to phraseology. A

[79] Thanks to Viktor Elšík for pointing this out.

major difficulty posed by phraseological sequences is the fact that they are often opaque, conventionalised, or do not lend themselves to direct translation. The first step towards addressing this topic in the language classroom may be a contrastive analysis of frequent English patterns and their equivalents in the students´ L1, which may subsequently inform teaching materials. This approach may be especially fruitful in EAP or ESP teaching, where corpora representing a particular relevant register could be used. It has been suggested e.g. by Biber (2006) and Biber et al. (2004), as well as indicated by the present study, that n-gram methodology seems efficient in revealing register-specific patterns, conveying functions associated with the particular register.

7 References and sources

Aarts, B., Chalker, S., & Weiner, E. (1994). *The Oxford Dictionary of English Grammar*. OUP.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101–122). OUP.

Anthony, L. (2017). *AntGram* (0.0.3) [Computer software]. Waseda University. http://www.laurenceanthony.net

Biber, D. (2006). *University Language. A Corpus-based Study of University Registers.* John Benjamins.

Biber, D., & Conrad, S. (2009). *Register, Genre and Style*. CUP.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman. https://books.google.co.uk/books/about/Longman_Grammar_of_Spoken_and_Written_En.html?id=vjomAQAAMAAJ&redir_esc=y

Biber, D., & Reppen, R. (Eds.). (2015). *The Cambridge Handbook of English Corpus Linguistics*. CUP.

Bondi, M. (2010). Perspectives on keywords and keyness. An introduction. In M. Bondi & M. Scott (Eds.), *Keyness in Texts* (pp. 1–20). John Benjamins.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. CUP.

Čermák, F. (2017). *Frazeologie a idiomatika česká a obecná. Czech and General Phraseology*. Karolinum.

Čermáková, A., & Chlumská, L. (2016). Jazyk dětské literatury: Kontrastivní srovnání angličtiny a češtiny. In A. Čermáková, L. Chlumská, & M. Malá (Eds.), *Jazykové paralely* (pp. 162–187). NLN.

Čermáková, A., & Chlumská, L. (2017). Expressing place in children's literature. Testing the limits of the n-gram method in contrastive linguistics. In T. Egan & H. Dirdal (Eds.), *Cross-linguistic*

Correspondences: From Lexis to Genre. John Benjamins. https://www.academia.edu/38144441/Expressing_place_in_children_Cermakova_Chlumska.pdf

Český národní korpus. (2016). *Srovnávací frekvenční seznamy (Reference frequency lists)*. Institute of the Czech National Corpus, Faculty of Arts, Charles University. https://wiki.korpus.cz/doku.php/seznamy:srovnavaci_seznamy

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, *11*(4), 411–433.

Colson, J.-P. (2008). Cross-linguistic phraseological studies. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 191–206). John Benjamins.

Coulthard, M. (2004). Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics*, *25*(4), 431–447. https://doi.org/10.1093/applin/25.4.431

Cowie, A. P. (1998a). Phraseological Dictionaries: Some East-West Comparisons. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications*. OUP.

Cowie, A. P. (1998b). *Phraseology: Theory, analysis and applications*. OUP.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Education Limited.

Cvrček, V. (2017). Zipfovy zákony [Zipf´s laws]. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *CzechEncy—Nový encyklopedický slovník češtiny*. Masarykova univerzita. https://www.czechency.org/slovnik/ZIPFOVY%20Z%C3%81KONY

Cvrček, V. (2021). *Calc: Korpusová kalkulačka [Calc: Corpus Calculator]* (1.02) [Computer software]. Czech National Corpus. http://www.korpus.cz/calc

Cvrček, V., & al. (2015). *Mluvnice současné češtiny 1: Jak se píše a jak se mluví* (2.). Charles University, Karolinum.

Cvrček, V., & Richterová, O. (Eds.). (2019). *Pojmy:asociacni_miry", Příručka*

ČNK. http://wiki.korpus.cz/doku.php?id=pojmy:asociacni_miry&rev=1554815423

Cvrček, V., & Václavík, J. (2015). Jednoznačnost a kontext. Kvantitativní studie [Unambiguity and context. A Quantitative study]. *Korpus Gramatika Axiologie*, *11*, 28–41.

de Saussure, F., Baskin, W., Meisel, P., & Saussy, H. (2011). *Course in General Linguistics*. Columbia University Press. https://books.google.cz/books?id=n6VFhwfLs0gC

Dovalil, V. (2012). Nad Dolníkovou Teorií spisovného jazyka o konceptuálních a metodologických problémech ve výzkumu jazykových norem a spisovné variety. *Slovo a Slovesnost*, *73*, 135–146.

Duguid, A., & Anna Marchi, John Morley, Charlotte Taylor, Alan Partington, Caroline Clark. (2005). *SiBol: Siena—Bologna English language newspapers corpus*. University of Siena, University of Bologna. https://www.sketchengine.eu/sibol-corpus/

Dušková, L. (2015). *From syntax to text: The janus face of functional sentence perspective*. Karolinum.

Ebeling, J., & Ebeling, S. O. (2013). *Patterns in Contrast*. John Benjamins.

Ebeling, S. O., & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 207–230). CUP.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language*, *64*(3), 501–538. https://doi.org/10.2307/414531

Fletcher, W. H. (2002). *KfNgram*. MD: USNA. kwicfinder.com/kfNgram/kfNgramHelp.html

Fried, M. (2013). Pojem konstrukce v konstrukční gramatice. *Časopis pro Moderní Filologii*, *95*(1), 9–27.

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast*, *14*(1), 58–72. https://doi.org/10.1075/lic.14.1.04gra

Granger, S. (2017). Academic phraseology. A key ingredient in successful L2 academic literacy. In H. Fjeld & J. Henriksen (Eds.), *Academic Language in a Nordic Setting – Linguistic and Educational Perspectives* (Vol. 9, pp. 9-27.). Olsen & Prentice.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. Advanced nonnative writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *52*(3), 229–252.

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology. An interdisciplinary perspective* (Vol. 139). John Benjamins Publishing Company. https://benjamins.com/catalog/z.139

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 27–49). John Benjamins.

Gray, B., & Biber, D. (2015). Phraseology. In *The Cambridge Handbook of English Corpus Linguistics* (pp. 125–145). CUP.

Gries, S. Th., Newman, J., & Shaoul, C. (2011). N-grams and the clustering of registers. *ELR Journal*, *5*. http://ejournals.org.uk/ELR/article/2011/1

Groom, N. (2010). Closed-class keywords and corpus-driven discourse analysis. In M. Bondi & M. Scott (Eds.), *Keyness in Texts* (pp. 59–78). John Benjamins.

Groom, N. (2017, November 1). *Phraseology: A Critical Reassessment* [Lecture].

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Hasselgård, H. (2017). Temporal expressions in English and Norwegian. In M. Janebová, E. Lapshinova-Koltunski, & M. Martinková (Eds.), *Contrasting English and other Languages through Corpora* (pp. 75–101). Cambridge Scholars Publishing.

Hasselgård, H. (2019). Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In V. Wiegand & M. Mahlberg (Eds.), *Corpus Linguistics, Context and Culture* (pp. 339–362). De Gruyter. 10.1515/9783110489071-013

Hirschová, M. (2017). In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *CzechEncy—Nový encyklopedický slovník češtiny*. https://www.czechency.org/slovnik/DEIXE

Hoey, M. (2005). *Lexical Priming. A new theory of words and language*. Routledge.

House of Commons Official Report. (2015). *597*(17). https://hansard.parliament.uk/pdf/commons/2015-06-16

Howarth, P. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics*, *19*(1), 24–44. https://doi.org/10.1093/applin/19.1.24

Hunston, S. (2008). Starting with the small words. *Patterns, Meaningful Units and Specialized Discourses. International Journal of Corpus Linguistics*, *13*(3), 271–295.

Hunston, S., & Francis, G. (2000). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. John Benjamins. https://doi.org/10.1075/scl.4

Hunt, P. (Ed.). (2005). *Children's literature: The development of criticism*. https://trove.nla.gov.au/work/16435403?q&versionId=45411178

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*, 4–21.

Johansson, S. (2007). *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies* (Issue v. 26). John Benjamins Publishing Co; eBook Collection (EBSCOhost). https://search.ebscohost.com/login.aspx?authtype=shib&custid=s1240919&profile=eds

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on.

Lexicography, *1*, 7–36.

Kjellmer, G. (1991). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics. Studies in Honour of Jan Svartvik* (pp. 111–127). Longman.

Klégr, A., & Malá, M. (2009). English Equivalents of the Most Frequent Czech Prepositions A Contrastive Corpus-based Study. *Proceedings of the Corpus Linguistics Conference, CL 2009, Conference in Liverpool, 20-23 July 2009*. http://ucrel.lancs.ac.uk/publications/cl2009/

Knowles, M., & Malmkjaer, K. (1996). *Language and Control in Children's Literature*. Routledge.

Křen, M. (2009). The SYN Concept: Towards One-Billion Corpus of Czech. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference*.

Křen, M., & Bartoň, T. – Hnátková, M. – Jelínek, T. – Petkevič, V. – Procházka, P. – Skoumalová, H.: (2010). *SYN2009PUB: korpus psané publicistiky*. Ústav Českého národního korpusu FF UK. http://www.korpus.cz

Lindquist, H., & Levin, M. (2008). Foot and Mouth: The phrasal patterns of two frequent nouns. In S. Granger & F. Meunier (Eds.), *Phraseology. An Interdisciplinary Perspective* (pp. 143–158). John Benjamins. https://benjamins.com/catalog/z.139.15lin

Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. Routledge.

Malá, M. (2019). English and Czech children's literature: A contrastive corpus-driven phraseological approach. In I. Headlandová Kalischová & M. Němec (Eds.), *Functional Plurality of Language in Contextualised Discourse. Eighth Brno Conference on Linguistics Studies in English. Conference Proceedings. Brno, 12–13 September 2019* (pp. 109–125). Masarykova univerzita. https://doi.org/10.5817/CZ.MUNI.P210-9767-2020-8

Malá, M., Šebestová, D., & Milička, J. (2021). The expression of time in English and Czech children's literature. In A. Čermáková, T. Egan,

H. Hasselgård, & S. Rørvik (Eds.), *Time in Languages*. Benjamins.

Milička, J. (2019). *Engrammer*. Institute of the Czech National Corpus, Faculty of Arts, Charles University. http://www.milicka.cz/en/engrammer/

Nesselhauf, N. (2004). What are collocations? In D. J. Allerton, N. Nesselhauf, & P. Skandera (Eds.), *Phraseological Units: Basic concepts and their application* (pp. 1–22). Schwabe Verlag.

NFNZ, N. fond nezávislé žurnalistiky. (n.d.). *Mapa médií*. Retrieved 5 January 2021, from http://www.mapamedii.cz/mapa/typologie/index.php

Nikolajeva, M. (2004). Narrative theory and children's literature. In P. Hunt (Ed.), *International Companion Encyclopedia of Children's Literature* (2nd, volume I ed., pp. 166–178). Routledge.

Nikolajeva, M. (2005). *Aesthetic Approaches to Children's Literature: An Introduction*. Scarecrow Press.

Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Journal Language Assessment Quarterly*, *15*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421

Paquot, M., & Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. https://doi.org/10.1017/S0267190512000098

Partington, A. (2004). 'Utterly content in each other's company': Semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, *9*(1), 131–156. https://doi.org/10.1075/ijcl.9.1.07par

Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora*, *5*(2), 83–108.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication*. Routledge.

https://doi.org/10.4324/9781315836027

Pinsent, P. (2014). Theories of Genre and Gender: Change and Continuity in the School Story. In C. Butler & K. Reynolds (Eds.), *Modern Children's Literature: An Introduction* (2nd ed., pp. 105–120). Macmillan International Higher Education. 10.1007/978-0-230-21149-0_11

Pražák, A., & Šmídl, L. (2012). *CzechParl: Czech Parliament Meetings*. LINDAT/CLARIN Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.

Rosen, A., Vavřín, M., & Zasina, A. J. (2020). *The InterCorp Corpus, version 13 of 1 November 2020*. Institute of the Czech National Corpus, Charles University. https://kontext.korpus.cz/

Sabban, A. (2008). Critical observations on the culture-boundness of phraseology. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 229–241).

Sainsbury, L. (2014). Chronotopes and Heritage: Time and Memory in Contemporary Children's Literature. In C. Butler & K. Reynolds (Eds.), *Modern Children's Literature: An Introduction* (2nd ed., pp. 187–201). Macmillan International Higher Education. 10.1007/978-0-230-21149-0_11

Scott, M., & Tribble, C. (2006). What counts in current journalism. Keywords in newspaper reporting. In *Textual Patterns: Key Words and Corpus Analysis in Language Education* (pp. 161–177). John Benjamins.

Šebestová, D., & Malá, M. (2019). Expressing Time in English and Czech Childrens Literature: A Contrastive N-gram-Based Study of Typologically Distant Languages. In *Language Use and Linguistic Structure: Proceedings of the Olomouc Linguistics Colloquium 2018* (pp. 469–483). Palacky University.

Šebestová, D., Malá, M., & Milička, J. (2019, June 1). *The expression of time* *in English and Czech children's literature: A contrastive phraseological perspective*. ICAME 40, Université de Neuchatel.

Silvennoinen, O. O. (2017). Not only apples but also oranges: Contrastive negation and register. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*, *19*. https://varieng.helsinki.fi/series/volumes/19/silvennoinen/

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. OUP.

Sinclair, J. (2004). *Trust the Text*. Routledge.

Sinclair, J., Jones, S., & Daley, R. (2004). *English Collocation Studies: The OSTI Report* (R. Krishnamurthy, Ed.). Bloomsbury Academic.

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

Stephens, J. (2005). Analysing texts: Linguistics and stylistics. In P. Hunt (Ed.), *Understanding children's literature: Key essays from the second edition of the International Companion Encyclopedia of Children's Literature* (2nd ed., pp. 73–85). Routledge.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, *7*(2), 215–244.

Thompson, P., & Sealey, A. (2007). Through children's eyes? Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics*, *12*(1), 1–23.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1/2), 61–82.

Uhlířová, L., & Kučerová, I. (2017). Slovosled. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *CzechEncy—Nový encyklopedický slovník češtiny*. Masarykova univerzita. https://www.czechency.org/slovnik/SLOVOSLED

V. Červená, J. Filipec, F. Havlová, M. Churavý, L. Janský, K. Kozlová, L. Kroupová, J. Machač, H. Marešová, V. Mejstřík, E. Michálek, B. Papírníková, E. Pokorná, B. Poštolková, M. Roudný, Z. Sochová, N. Svozilová, E. Vodrážková, J. Zima. (2011). In B. Havránek, J. Bělič, M.

Helcl, & A. Jedlička (Eds.), *Slovník spisovného jazyka českého (online)*. Ústav pro jazyk český AV ČR (Institute of the Czech Language, Czech Academy of Sciences). https://ssjc.ujc.cas.cz/search.php?db=ssjc

Vašků, K., Brůhová, G., & Šebestová, D. (2019). Phraseological Sequences Ending in of in L2 Novice Academic Writing. In G. Corpas Pastor & R. Mitkov (Eds.), *Computational and Corpus-Based Phraseology. EUROPHRAS 2019* (pp. 431–443). Springer. https://doi.org/10.1007/978-3-030-30135-4_31

Vavřín, M., & Rosen, A. (2015). *Treq (v. 2.1)*. Treq. https://treq.korpus.cz/

Verhagen, A. (2015). Grammar and Cooperative Communication. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics* (pp. 232–252). De Gruyter Mouton.

Webb, S. (2019). *The Routledge Handbook of Vocabulary Studies*. Routledge.

Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, *22*(2), 212–241. https://doi.org/10.1075/ijcl.22.2.03wri