Lang Resources & Evaluation (2020) 54:713–745
https://doi.org/10.1007/s10579-020-09487-4

ORIGINAL PAPER

Comparing web-crawled and traditional corpora

Václav Cvrček[3] · Zuzana Komrsková[3] · David Lukeš[3] · Petra Poukarová[3] · Anna Řehořková[3] · Adrian Jan Zasina[3] · Vladimír Benko[1,2]

Published online: 19 March 2020

Abstract Using a multi-dimensional (MD) analysis of register variability, the study compares two corpora of Czech: Koditex, a "traditional" corpus carefully designed using various sources with rich metadata, and Araneum Bohemicum Maximum, a web-crawled corpus with an opportunistic composition representative of the "searchable" web. Both types of corpora are projected onto the space induced by the MD model, with the main objective being to find out whether they overlap in the linguistic variation they cover, or whether one introduces some specific variation which cannot be found in the other. We also document a crucial methodological

✉ Václav Cvrček
vaclav.cvrcek@ff.cuni.cz

Zuzana Komrsková
zuzana.komrskova@ff.cuni.cz

David Lukeš
david.lukes@ff.cuni.cz

Petra Poukarová
petra.poukarova@ff.cuni.cz

Anna Řehořková
anna.rehorkova@ff.cuni.cz

Adrian Jan Zasina
adrian.zasina@ff.cuni.cz

Vladimír Benko
vladimir.benko@juls.savba.sk

[1] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

[2] UNESCO Chair in Plurilingual and Multicultural Communication, Comenius University in Bratislava, Bratislava, Slovakia

[3] Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

point which has broader relevance for MD analyses in general, namely that texts have to be of similar lengths in order for their scores on the dimensions to be comparable. Results indicate that some traditional text categories, such as journalism or non-fiction, are characterized by language phenomena which are equally well covered by web-crawled data, though of course traditional corpora keep their edge in terms of the richness of the accompanying metadata. But overall, the range of variation in Koditex is broader as it contains texts which have no adequate substitute (i.e. texts with a comparable set of linguistic characteristics, regardless of their extratextual label) in data acquired through general-purpose web-crawling techniques. These include informal conversations, private correspondence, some types of fiction, but also user-generated content (comments on Facebook, forums etc.).

Keywords Web corpus · Crawling · Register · Variation · Multi-dimensional analysis · Czech

1 Introduction

1.1 Motivation

The last decade and a half in corpus linguistics has been heavily influenced by the advent of web-crawled corpora and the web-as-corpus approach. The popularity of web-crawled corpora rests on several factors, the most obvious ones being unrivaled size (cf. the 14 billion word iWeb corpus of English, Davies 2018) and low cost. Linguists longing for larger data were supplied with unprecedented volumes of text crawled from the web without the need to undergo the tedious and expensive process of text collection and sorting. Furthermore, using the same web-crawling method, a set of comparable corpora for several languages can be compiled; thus we have the WaC family (Baroni et al. 2009), the TenTen family (Jakubíček et al. 2013) or the Aranea family (Benko 2014), to name at least a few of them. The appeal of crawling also lies in the fact that with currently available tools (Anthony 2018; Baroni et al. 2006), it does not require expert IT knowledge and the copyright restrictions are in many cases less strict.

If this is the case, should we still invest into building more expensive traditional corpora? Crawling is cheaper chiefly because it is opportunistic—it gathers whatever is easily accessible on the web and lumps it together, leaving it up to the user to sort it all out. By contrast, traditional corpora are distinguished by careful design, determining well-motivated text categories, assigning quotas and sticking to them, even though the data might be hard to acquire. Let us emphasize that the source of the data is an orthogonal issue—traditional corpora may very well contain print, handwritten, spoken or web data—the important thing is the design effort.[1]

[1] We consider the following projects as prototypes of traditional corpora: BNC for English (Aston and Burnard 1998) or NKJP for Polish (Górski and Łaziński 2012), each containing a majority of written texts and some proportion of spoken and/or web communication.

One consequence of the traditional approach is that it yields text category metadata which on its own is worth the higher cost if you need it (and many types of research rely on it). On the flip side, it can also lead you down the rabbit hole of chasing hard to acquire data which might ultimately turn out to be completely interchangeable, from the point of view of the linguistic features it exhibits, with a different, more convenient source. So the question becomes, given the overall variation in a language, are there substantial areas which are not represented by easily accessible data from the web? If so, what are they? And conversely, are there perhaps areas which are only revealed if large enough data are collected, which is what general opportunistic web-crawled corpora excel at?

As an example, take the representation of spoken language. Some linguists believe that some electronic texts found on the web share key characteristics with spoken conversation (Baron 2010; Herring 2010; specifically for Czech cf. Hoffmannová et al. 2016, p. 105). It might thus seem reasonable to use a web-crawled corpus as a representation of both oral and literate discourse—but is it truly, and if so, to what extent and in what sense?

Such questions can be answered by comparing web-crawled and traditional corpora with respect to the amount of linguistic variation they cover. This has been a concern for as long as web-crawled corpora have existed, and some early work in this field, due to Ide et al., has expressed serious reservations in this regard: "It is not at all clear that a web corpus can be balanced for genre, and it is likely that certain genres, such as fiction will be under-represented" (Ide et al. 2002, p. 840). The study concludes that "web-based texts are, in any case, representative of only a small slice of the range of genres encountered by human readers everyday, and therefore cannot be used to provide a comprehensive view of American English in the 1990's" (Ide et al. 2002, p. 844).

However, since Ide et al. wanted to focus on American English, they only crawled institutional *.gov* and *.edu* domains, which constitutes a noticeably biased sample of the writing openly available on the web. Also, the web has become even more ubiquitous and all-encompassing in the intervening years, so it is possible the situation has changed to a certain extent. Finally, as the lingua franca of the web, English is an outlier and cannot be used as a yardstick for other languages. For all these reasons, we believe a fresh perspective on the subject to be useful, and propose one centered around the Czech language community, which has a vibrant and active web presence, though of course nowhere near that of English.

1.2 Theoretical background

The issue of representativeness—which is at the core of this research—has been addressed several times (e.g. Leech 2007; Cvrček et al. 2016). Usually, it refers to the relation a particular corpus, as a sample of texts, bears to the population of all texts in a language. However, there is no way to quantify such a relation, as we cannot access this entire population. Consequently, this paper attempts to bootstrap a quantifiable examination of representativeness by comparing the ranges of variation covered by two competing corpora. Our starting point is Biber's definition of representativeness, accentuating coverage of language phenomena:

"Representativeness refers to the extent to which a sample includes the full range of variability in a population" (Biber 1993, p. 243). What does this mean in practice in the context of corpus design? In assessing representativeness, both situational and linguistic perspectives are important: "Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language" (Biber 1993, p. 243).

The first criterion presupposes availability of reliable text category metadata, which is overwhelmingly missing in web-crawled corpora. In contrast to traditional corpora, web-crawled ones usually contain only "technical" metadata derived from the circumstances in which the text was obtained, such as domain, URL, time, size etc. Compare with the amount of metadata included in the traditional corpus of written Czech SYN2015, in which text classification alone spans a four-level hierarchy (Křen et al. 2016). Some interesting attempts at automatic text-type or register annotation (Sharoff 2018) and analyzing internet-specific registers (Biber and Egbert 2016) have been made for English; however, a broader perspective is missing that would establish parallels between online and offline registers, enabling direct comparison of the variability they cover, and the situation in other languages, which may show cultural specificities, is virtually uncharted. In the absence of metadata, we decided to focus on the second criterion mentioned by Biber, linguistic properties of texts.

The topic of corpus comparison keeps cropping up in the literature (Rayson and Garside 2000; Kilgarriff 2001, 2012; Piperski 2017, 2018), with the majority of these approaches using frequency distributions of words or other units, or a combination of lexical and grammatical features (Sharoff 2018). Our approach differs in that it embraces multi-dimensional analysis (MDA; Biber 1988, 1995), incidentally much like Ide et al. (2002): instead of relying on isolated features, we see them as entangled in complex relationships with other features, forming dimensions of variation. It should be noted that MDA was devised as a tool for assessing functional/text-linguistic/register variation, so this is what the comparisons reflect. Comparisons articulated around other types of variation, e.g. topic-related or sociolinguistic, can conceivably yield different results.

Note that in the following, we refer to different text categories using extratextual labels (e.g. private correspondence), but the comparison is based purely on intratextual features. We are thus not trying to compare lists of text categories occurring in the metadata of traditional vs. web-crawled corpora—it would be trivial to check whether private correspondence appears in both of them. What MDA can do is to help us ascertain whether web-crawled data contains any texts which even *read like* private correspondence, irrespective of their extratextual labels.

As a related question, we also tackle the issue of text length (or the length of text excerpts in our case) and the impact it has on the results of MDA and on comparing the variation covered by corpora in general. Since we are dealing with text excerpts, we are not concerned with text length as a characteristic property of some genres (e.g. novels are usually longer than stories), but simply with how text length interacts with the mathematics of the statistical procedure of factor analysis, which is at the heart of MDA. Our goal is to test a hypothesis articulated in previous

research comparing two corpora with unequal excerpt lengths via MDA: "since [the examined] texts are shorter than those [the MD model is built on] by an order of magnitude, their dispersion in dimensions is considerably higher (…) it is a natural consequence of the fact that shorter texts are inevitably more homogeneous, more distinctive register-wise and therefore more likely to incline towards dimension extremes" (Cvrček et al. forthcoming). If this is confirmed, then text length should be controlled for in MDA-based corpus or register comparisons, to eliminate its potential confounding influence on the results.

1.3 Overview of the study

The outline of the paper closely follows the course of actions carried out in this research project, which can be summarised in the following steps:

1. First, we compiled Koditex, a traditional corpus consisting of excerpts covering a wide range of available texts in Czech (see Sect. 2.1).
2. A list of 122 features which reflect the functional variability of texts was assembled and operationalized.
3. The features were identified in Koditex texts and their values (mostly relative frequencies) were submitted to factor analysis in order to establish dimensions of variation (i.e. the MD model); based on the loadings of features, individual dimensions were interpreted and labeled (see Sect. 2.2).
4. The general opportunistic web-crawled corpus Araneum Bohemicum, representing the Czech searchable web, is sampled for subcorpora which will be used for the comparison. The source corpus is introduced in Sects. 3.1 and 3.2, the sampling procedure is described in Sect. 3.3.
5. The set of 122 linguistic features assembled for the original MD model is applied to these web sample texts using the same operationalizations; Sect. 4 discusses some of the methodological issues around this.
6. Based on the feature values obtained in the previous step, the positions of the web sample texts within the variation space induced by the MD model in step 3 are calculated (or, in other words, texts from the web samples are projected onto the dimensions of the MD model by calculating their factor scores).
7. At this point, we know the coordinates in MD space of the texts both from the traditional Koditex and the web-crawled corpus samples, so we can compare the ranges of variation covered by these two types of corpora, identifying overlaps and Koditex- versus web sample-specific complements on different dimensions (see Sects. 5.1 and 5.3). Where expedient and possible, we describe these overlaps and complements in terms of extratextual characteristics (e.g. genres) for the sake of conciseness and clarity: for instance, a statement such as "public speeches are fully covered by web data" is meant as shorthand for "the region of MD space where public speeches are typically encountered is well covered by web data"; whether the web data actually contains public speeches or some other genres which just happen to be linguistically equivalent is beyond the scope of this paper.

8. The related issue of the influence of text length on MDA results is examined in Sect. 5.2.

Steps 1–3 were performed as part of previous research on the MDA of Czech and reported in full detail in Cvrček et al. (2018a, b). Nevertheless, an extensive summary is provided in Sect. 2 for the reader's convenience, so as to make subsequent comparisons based on the MD model more tangible.

2 The MD model derived from a traditional corpus

2.1 The Koditex corpus

The multi-dimensional model of Czech register variation (Cvrček et al. 2018a, b) is based on Koditex (Zasina et al. 2018), a 9-million-word corpus (10.9 million tokens incl. punctuation) designed to be as representative of the wide range of uses of language as possible, following purely extratextual criteria (i.e. derived from available metadata). Table 1 gives an overview of the corpus at different levels of granularity; the topmost category, mode of communication (written, spoken, web), is subdivided into 8 divisions, which ultimately further differentiate into 45 classes of texts, aiming at roughly 200,000 words per class, subject to data availability. A detailed overview of the corpus, including data sources and annotation tools employed, is available at https://wiki.korpus.cz/doku.php/en:cnk:koditex or in Zasina and Komrsková (2019).

Given the focus of this paper, a specific note on the web mode is necessary. This part of the Koditex corpus was not designed as an opportunistic sample of texts which can be found on the internet, on the contrary, it was conceived as a representation of web-specific genres (i.e. genres which cannot be found outside the web).[2] Therefore a careful selection of domains was employed in order to sample the 5 classes established within the *web* mode. Three classes are grouped into the "multi-directional" division, which covers situations where multiple parties interact: public Facebook statuses, discussion forums and comments sections. The remaining two, blogs and Wikipedia articles (which might differ from traditional encyclopedias simply due to the absence of space limitations), form the "uni-directional" division, where the roles of reader and writer are more clearly separated. We do not claim this to be an exhaustive list, e.g. instant messaging could have been another category of interest, had we had the data. As it stands, the *web* mode part of the Koditex corpus was assembled using the following sources:

[2] The reason why we selected web texts for Koditex based on broad genres derived from extratextual metadata, as opposed to more fine-grained registers (as in Biber and Egbert 2016), is twofold: (1) a typology of web registers in Czech has not been established to date, and (2) we did not have any information about the web texts in the process of Koditex compilation other than their source, we therefore could not classify them according to their register prior to the MDA being carried out.

Table 1 Overview of the structure of the Koditex corpus

Mode	Division	Superclass	Class	Tokens	Text chunks
spo (spoken)	int (interactive)		bru (unprepared broadcast discussions)	221,812	90
			eli (elicited speech)	201,690	82
			inf (informal unprepared private dialogue)	208,565	86
	nin (non-interactive)		wbs (written-to-be-spoken speeches)	213,201	71
web	mul (multi-directional)		dis (discussions)	197,948	87
			fcb (Facebook posts)	199,418	91
			for (forums)	200,104	85
	uni (uni-directional)		blo (blogs)	204,356	74
			wik (cs.wikipedia.org articles)	201,691	84
wri (written)	fic (fiction)	nov (novels)	crm (crime)	190,026	68
			fan (fantasy)	189,432	69
			gen (general fiction)	193,667	67
			lov (romance)	189,893	70
			scf (sci-fi)	188,703	68
			col (short stories)	195,595	70
			scr (screenplays & drama)	182,689	76
			ver (poetry & lyrics)	205,837	76
	nfc (non-fiction)	pop (popular science)	fts (formal and technical sciences)	207,607	68
			hum (humanities)	204,837	74
			nat (natural sciences)	204,751	71
			ssc (social sciences)	203,698	68
		pro (professional journals)	fts (formal and technical sciences)	210,010	71
			hum (humanities)	207,916	69

Table 1 continued

Mode	Division	Superclass	Class	Tokens	Text chunks
			nat (natural sciences)	209,580	70
			ssc (social sciences)	209,385	72
		sci (scientific/academic)	fts (formal and technical sciences)	202,932	67
			hum (humanities)	204,300	71
			nat (natural sciences)	206,716	72
			ssc (social sciences)	205,358	67
			adm (administrative texts)	203,542	82
			enc (encyclopedias)	203,957	73
			mem (memoirs)	203,390	71
	nmg (newspapers and magazines)	lei (leisure)	hou (crafts and hobbies)	207,499	68
			int (interesting facts)	209,232	69
			lif (lifestyle)	203,124	72
			mix (supplements, Sunday magazines)	205,310	75
			sct (tabloids)	201,417	73
			spo (sport)	199,238	70
		new (newspapers)	com (op-eds, columns)	205,372	68
			cul (culture)	205,690	68
			eco (economic news)	211,481	70
			fre (free time activities)	208,532	71
			pol (politics)	206,893	70
			rep (news)	206,377	70
	pri (private)		cor (letters)	96,366	68
Total				9,039,137	3292

Token counts exclude punctuation

- blogs were extracted from the Araneum Bohemicum corpus (for details see 3.1; we made (we made sure texts included in Koditex were excluded from the web sample corpora used for comparison in this study).
- posts from Facebook, forums and comments sections (the *fcb*, *for* and *dis* classes) were sampled from data which was kindly provided by Josef Šlerka and the team at SocialInsider, a then-leading tool for monitoring user-generated content on the Czech internet.
- the *wik* class was sampled from a corpus of Czech Wikipedia articles compiled at NLPC in Brno (https://nlp.fi.muni.cz/en/NLPCentre).

Even though there might be some overlap with the web-crawled data to which the Koditex-based model will be compared, the scope, purpose and focus of the web representation is fundamentally different. In this comparison, Koditex occupies the "designed and traditional" end of the spectrum, to which the opportunistic approach of a catch-all web-crawled corpus is compared.

For reasons that will become clear after presenting the comparison results in Sect. 5.3, a short note on the composition of the spoken interactive division of Koditex is necessary. Texts in these classes come from various corpora of spoken language:

- ORAL2013 (Válková et al. 2012; Benešová et al. 2013), a corpus of unprepared, private dialogues (between family and friends) collected across the entire Czech Republic, used for populating the *inf* class.
- PMK—Prague spoken corpus (Čermák et al. 2001) and BMK—Brno spoken corpus (Hladká 2002)—the "formal" parts of these corpora, which consist of elicited speech (semi-structured interviews), were used for the *eli* class.
- DIALOG (Kaderka 2012)—a multimedia corpus of unprepared broadcast speech containing transcripts of a fairly wide range of discussion programs broadcast on Czech TV stations; used for the *bru* class.

To ensure higher variability, each class consists of continuous text samples referred to as "chunks", rather than entire texts.[3] Chunking allows to control for the observation units' length (details and reasoning are described in Cvrček et al. 2018b). Where necessary, longer texts were split and shorter aggregated in order to achieve homogeneous chunks of 2000–5000 running words, with some exceptions going as low as 1000 where data was scarce.

2.2 Dimensions of variation

Following the MDA methodology devised by Biber (1988), we assembled a list of 122 linguistic features related to functional variability in Czech. The list was inspired by Biber's original set and supplemented with features of Czech known to participate in variation according to secondary literature. Features were

[3] A similar motivation for using text excerpts can be found in the Brown corpus (Francis and Kučera 1964).

operationalized, their values for the Koditex chunks were retrieved, and the data was statistically evaluated by factor analysis using the *fa* function from the R *psych* package (R Core Team 2018; Revelle 2018), with *promax* rotation.[4] Using heuristics described in Cvrček et al. (2018b), we settled on a model with 8 dimensions (labeled GLS1–GLS8 by the factoring method) which were then interpreted and assigned descriptive labels for their positive and negative poles.

The interpretation is based on (1) the loadings of linguistic features on the factors (i.e. the extent to which the features push a text towards the extremes of a given dimension) and (2) the factor scores of Koditex chunks (i.e. coordinates of a text on individual dimensions). A quick overview of this interpretation is given below; for a more detailed description, the reader is referred to previous work (Cvrček et al. 2018b, Sect. 5).

2.2.1 Dimension 1: dynamic (+) vs. static (−)

This dimension explains the largest proportion of variability (21.5%, see Table 2). The positive extreme is strongly associated with features concerning verbal categories: past tense, verbs in general, finite verbs, and the indicative. On the negative side, the most prominent features are nominal post- and pre-modifiers, adjectives, abstract nouns, and clusters of nouns and adjectives. Analogous feature groupings were obtained in previous MD analyses; for instance, Biber (2014) describes this dimension as clausal vs. phrasal discourse, reflecting whether the primary text building strategy is clause chaining or inner elaboration of individual phrases. With regard to genre, this dimension separates novels, private correspondence and web forums on the positive pole from administrative documents, natural sciences, formal and technical texts, encyclopedias and Wikipedia entries as extreme instances on the negative pole.

2.2.2 Dimension 2: spontaneous (+) vs. prepared (−)

The positive end shows features characteristic of spontaneous texts: contact expressions, fillers, demonstrative pronouns, and interjections. On the negative side, features such as nominal cases with prepositions, clauses with interrogative/relative adverbs, and prepositions (both frequent and varied) indicate more complex, elaborated discourse. The text chunks associated with the positive pole are represented by interactive spoken communication (private, elicited and broadcast), whereas the negative pole brings together prepared texts such as administrative documents, Wikipedia articles and scientific texts in technical domains.

2.2.3 Dimension 3: higher (+) vs. lower (−) level of cohesion

The main positive features include subordinating correlative connectives, predicative nouns, relative subordinate clauses, possessive pronouns, and use of a wide

[4] The data set is available via the TROLLing repository (doi: https://doi.org/10.18710/QAJKZW). It also includes the full list of linguistic features employed.

Table 2 Summary of the variance explained by the MD model

	GLS1	GLS2	GLS5	GLS8	GLS3	GLS7	GLS4	GLS6
Proportion of variance explained	0.215	0.142	0.044	0.039	0.034	0.032	0.032	0.022
Cumulative variance explained	0.215	0.357	0.401	0.439	0.474	0.505	0.537	0.559

The total amount (56%), is comparable to the 52% reported by Biber for English (Biber 1995, p. 121)

repertoire of pronouns. The texts with positive scores are represented by scripted speeches, social sciences and TV talk shows. On the other side, there appear encyclopedic entries, informal spoken communication and Wikipedia articles. To generalize, the third dimension captures the difference between just getting facts across (−) and structuring them into a coherent sequence (+).

2.2.4 Dimension 4: polythematic (+) vs. monothematic (−)

Among the features with positive loadings, we find bigram and unigram richness, toponyms, and use of a varied repertoire of prepositions and pronouns. Conversely, the following features came up with negative loadings: thematic concentration of text, lexical repetitiveness, verbal nouns, and passive voice. We interpret this as a contrast between poly- and monothematicity. The typical genres on the polythematic (+) side are arts and entertainment sections from newspapers, leisure magazines or tabloids, while on the monothematic (−) pole, they are administrative documents and scientific and professional technical texts.

2.2.5 Dimension 5: higher (+) vs. lower (−) amount of addressee coding

The common characteristic of the features with positive loadings is that they encode direct or indirect references to an addressee: questions, verbs in the second person, non-polar questions, and second person pronouns. The negative pole of this dimension only has one prominent feature, average sentence length, which indirectly signals a focus on content rather than interaction. In Koditex, low amounts of addressee coding can be found in elicited spoken communication (only answers were recorded, not the questions), Wikipedia articles and TV discussion programs, whereas high amounts are attested in screenplays, poems, science fiction, and some types of novels, particularly crime and fantasy.

2.2.6 Dimension 6: general (+) vs. particular (−)

The key to interpreting this dimension lies in the salient negatively loading features: proper names, numerals and time expressions. These all refer to concrete, particular information, while the positive pole is associated with more abstract linguistic devices such as coordination, semantically bleached adjectives, and use of a wide repertoire of conjunctions. The negative, particular extreme is exemplified by text categories such as sports and economic news and screenplays, whereas the positive, general extreme is associated with encyclopedias or poems.

2.2.7 Dimension 7: prospective (+) vs. retrospective (−)

Feature-wise, this dimension rests on an opposition between, on the one hand, present and future tense, predicative adjectives, the imperative, and second person verb forms (+), and on the other, past tense, third person pronouns, possessive adjectives and pronouns, and relative subordinate clauses (−). As for text chunks, while highly prospective ones come from elicited spoken communication, web forums and private correspondence, highly retrospective ones are found in novels, particularly crime, science fiction and fantasy. This suggests another helpful way of thinking about this opposition, one which previous MD studies have often adopted: non-narrative (+) vs. narrative (−).

2.2.8 Dimension 8: attitudinal (+) vs. factual (−)

Positively loading features include hedges and downtoners, restrictors, and intensifiers and boosters, along with adverbial expressions and use of a wide inventory of conjunctions. These are associated with expression of opinion, evaluation and openly acknowledged subjectivity. The negative extreme does not give us too many clues in terms of features, perhaps only a tendency to enumeration—the only salient feature is coordination. As for text chunks, web discussions, private correspondence and web forums come out as attitudinal, whereas administrative documents, Wikipedia articles and screenplays as factual.

When using this MD model to contextualize other data, e.g. web-crawled corpora, we should bear in mind that not all dimensions are equally important. The order in which the dimensions are listed above does not reflect the amount of variation each one accounts for. Table 2 summarizes the proportion of variance explained by each factor as well as showing the cumulative numbers.

3 Web-crawled data

The MD model based on the Koditex corpus will be contrasted with several samples of data from the web-crawled corpus Araneum Bohemicum, which was created within the framework of the Aranea Project (Benko 2014, 2016a). The present study is based on Araneum Bohemicum Maximum 15.04, a version compiled from data crawled during the initial phase of the project in early 2013.

3.1 The Aranea project

The aim of the Aranea Project is to create a family of very large (i.e., containing more than 1 billion tokens) web-based corpora that can be used as a resource for teaching languages and translation/contrastive studies. All corpora follow the Web as Corpus (WaC) methodology (Baroni et al. 2009) and are built using tools developed (mostly) at the Faculty of Informatics at the Masaryk University in Brno (Kilgarriff et al. 2010). The main components of the toolchain are: SpiderLing, a

specialized web crawler optimized for downloading textual data (Suchomel and Pomikálek 2012), Onion, a tool to deduplicate documents and/or their parts (Pomikálek 2011), and Unitok, a universal tokenization procedure (Michelfeit et al. 2014). As all corpora are processed by a unified pipeline (Benko 2016b) and have the same size, they can to some extent be considered "comparable" with respect to the way the data was collected and, in fact, they are often used for contrastive research.

Crawling for the Czech corpus was performed in several sessions during May and June 2013 and yielded 9.5 million documents (web pages) containing approx. 5.5 billion tokens of text. After filtering and deduplication, these statistics dropped to 5.2 million documents and 3.3 billion tokens, respectively.

3.2 Composition

Given the general lack of linguistically relevant metadata in web corpora mentioned in the introduction, it is hard to characterize the composition of Araneum Bohemicum from a register perspective. We can however get an idea of the composition of the corpus in terms of the sources of the data, i.e. the domains from which it was crawled. Two chief criteria come to mind: (1) coverage of relevant domains, and (2) distribution of data within domains.

Regarding the first parameter, it is important to note that crawling can only get data from the searchable and freely accessible segment of the web, i.e. that which is not hidden behind paywalls and other mechanisms requiring registration and logging in (unlike the *web* mode portion of the Koditex corpus described in Sect. 2.1, which covers also social media, forums etc.).[5] SpiderLing also obeys all restrictions defined in the *robots.txt* files for the respective web sites. Therefore, we can unfortunately only make a rather rough estimate of coverage. According to the statistics provided by the *cz.nic* domain administrator, in 2013, when the first version of Araneum Bohemicum was crawled, there were approx. 1.1 million registered second-level internet domains under the *.cz* national top-level domain.[6] Out of these, some 151,000 domains are present in the corpus, which is approx. 15%.

With respect to assessing the distribution of data within domains, the task is easier as we can simply analyze the collected data. Crucially, SpiderLing takes proactive measures for balancing the data set by means of a dynamic yield rate threshold formula to prevent downloading too many documents and/or too much data from a single domain (for details see Suchomel and Pomikálek 2012, Sect. 3.1). Binning the data by the 210,000 domains present in the Araneum Bohemicum Maximum corpus, we find that this built-in functionality works fairly well—even the most frequent domains do not reach a proportion larger than a small fraction of a percent.

[5] It should be pointed out that apart from general opportunistic web-crawled corpora such as Araneum Bohemicum, there are also specialized web corpora concentrating on specific domains (corpora of tweets etc.), typically requiring more targeted approaches to data collection, e.g. through provided custom APIs.

[6] Cf. https://stats.nic.cz/reports/2013/ (visited November 2019).

Based on the foregoing discussion, we believe that Araneum Bohemicum Maximum can be considered an inclusive and varied sample of the Czech searchable web in 2013, given the technology employed and the limits of the available technical infrastructure.

3.3 Sampling

The sampling method used for obtaining samples from Araneum Bohemicum followed the same principles that were used for Koditex: (1) text excerpts instead of entire texts,[7] (2) an emphasis on the maximum diversity of texts (taking into account what little available and relevant metadata there was), (3) text length control (chopping up longer texts into chunks). In order to improve the reliability of the analysis, three batches of samples from the Araneum Bohemicum corpus were used—WS-K1, WS-K2 and WS-S. Since one of the working hypotheses was that the dispersion of texts in the MD space may be influenced by the distribution of text lengths (because longer texts gravitate towards the center of the scale by virtue of their mixed composition), the three batches are of two types:

- two batches with 5000 text excerpts each and text length distributions modeled after Koditex (named WS-K1 and WS-K2—"web samples with Koditex distribution"). Two batches were made in order to minimize a possible sampling error which can occur especially when sampling relatively small samples from a large population. The WS-K1 and WS-K2 batches should not differ in any respect, therefore one should be considered as a control data set for the other.
- one batch with 1000 text excerpts and the same text length distribution pattern, but with texts shorter by an order of magnitude (henceforth WS-S—"web sample short").

We should stress that it is not the case that originally longer texts are more likely to end up in WS-K* batches, whereas shorter ones in WS-S. The original texts are assigned to batches randomly and then chunked in order to fit the requirements. In other words, there is no sampling bias between the WS-K* batches and WS-S.

Table 3 shows a comparison of the means and standard deviations of the batches with Koditex. The batches contain mutually exclusive text excerpts from different sources: they were crawled from different URLs and the WS-K1 and WS-K2 batches contain uniform proportions of beginnings, ends and middle portions of the original texts.

[7] As a matter of fact, in the domain of hypertext media, the concept of "entire text" is problematic anyway.

Table 3 Means and standard deviations of text lengths in the corpora used

Corpus	No. of text chunks	Mean	Std. dev.
Koditex	3292	2745.8	748.6
WS-K1	5000	2743.3	772.1
WS-K2	5000	2748.4	771.2
WS-S	1000	290.8	74.4

4 Comparison methodology

As outlined in Sect. 1.3, the procedure of analyzing WS data against the backdrop of the general-purpose MD model of Czech register variability derived from the Koditex corpus follows two steps:

1. Take the set of 122 linguistic features assembled for the original MD model and, using the same operationalizations, evaluate them on the chunks in the WS batches (the only exception being that due to the amount of data, there were no manual checks).[8]
2. Chunks from all batches were projected onto the dimensions of the MD model by calculating their factor scores (i.e. estimates of the chunk's position within each dimension). For each WS text with respect to each dimension, this means computing the sum of the factor weights of the original MD model multiplied by the standardized values of the corresponding features extracted from the text.

By comparing the factor scores of WS batches with those originally computed for Koditex when establishing the MD model, we can find out the degree of overlap between these two sources of information about linguistic variation in Czech (opportunistic web-crawled corpus vs. carefully designed traditional corpus). Before proceeding to the results, a few methodological remarks are in order.

First, the comparison of any model resulting from factor analysis and a new dataset may be influenced by a phenomenon called *factor scores indeterminacy*: "an infinite number of ways for scoring the individuals [i.e. text chunks in our case] on the factors could be derived that would be consistent with the same factor loadings" (Grice 2001, p. 431). While there is, in general, no guarantee that any factor score (either of the texts originally used to establish the MD model, or of any newly evaluated ones) represents the best fit (in fact there may be several acceptable solutions to the factor model other than the one chosen), there exist indicators of model indeterminacy and reliability. Grice (2001, p. 435) mentions the multiple correlation between each factor and the original variables and the minimum possible correlation between two sets of competing factor scores. In our case, the first index ranges from 0.98 to 1 for the individual dimensions and the second one is between 0.92 and 0.99. These values are high enough not to warrant any particular caution in interpreting the results.

[8] We did not take any special measures to deal with the noise in the web-crawled data as we wanted to keep the operationalizations identical.

Second, an important question is, how exhaustive is the comparison provided by the approach adopted here? Since we use MDA as a means for comparison, it is inevitably tied to the set of features included in the model. A truly comprehensive comparison of two corpora, on the other hand, should comprise the full set of grammatical as well as lexical features included in the corpora. This is obviously beyond the scope of this method, as MDA is designed to investigate functional variation only and in practice cannot take into consideration all grammatical and lexical features. However, given the large number of different linguistic features (122) from different levels (phonology, morphology, lexicon, syntax and pragmatics) used in our MDA, and taking into account that the model explains 56% of the variation in the data, we believe the comparison presents a sophisticated and informative picture of the scope of text linguistic variation covered by the two types of corpora.

As for the interpretation of the comparison, two perspectives are used in this study. In neither do we further reduce the dimensionality of the data, though statistical techniques such as t-distributed stochastic neighbor embedding (t-SNE; see Sharoff 2018) could be used for this purpose. The reason is that the 8 dimensions of our model are linguistically meaningful, and we want to describe the differences in relation to the register functions each of them carries. First, we present a per-dimension comparison, comparing the variance of Koditex and WS batches within each dimension separately. In this part of the comparison, we will pay attention not only to the extent of overlap between the Koditex and WS texts on each dimension, but also to the range of variation specific to each of the corpora (see Fig. 1).

Second, we present a comparison of the web-crawled corpora and the Koditex data in a simplified perspective, referring only to the two most important dimensions of the MD model (dynamic vs. static and spontaneous vs. prepared, cf. Table 2). These two dimensions cover most of the variance explained by the model (almost two-thirds) and therefore represent the most important sources of variation. As will be shown, the advantage of this approach is that data in 2 dimensions can be easily visualized (as opposed to 8 dimensions), which means that we can intuitively spot text categories which might not be covered by one of the data sources.

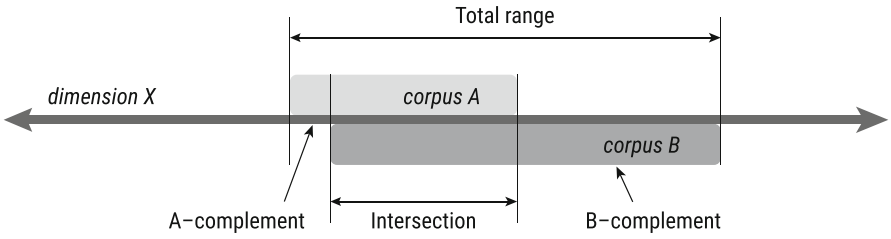

Fig. 1 Schematic representation of the variation ranges covered by corpora A and B on dimension X of an MD model

5 Results and interpretation

5.1 Per-dimension comparison

The spread of the chunks' factor scores within all eight dimensions is summarized by the boxplots in Fig. 2. The original MD model is represented by the texts of the Koditex corpus, the Araneum Bohemicum corpus is represented by two samples with text length distributions modeled according to Koditex (WS-K1 and WS-K2), and one sample containing texts which are ten times shorter (WS-S). While the position of the median does not seem to vary substantially, the dispersions differ, sometimes substantially (tests of homogeneity of variances—Bartlett, Fligner-Killeen—show that the differences are significant, $p < 0.01$).

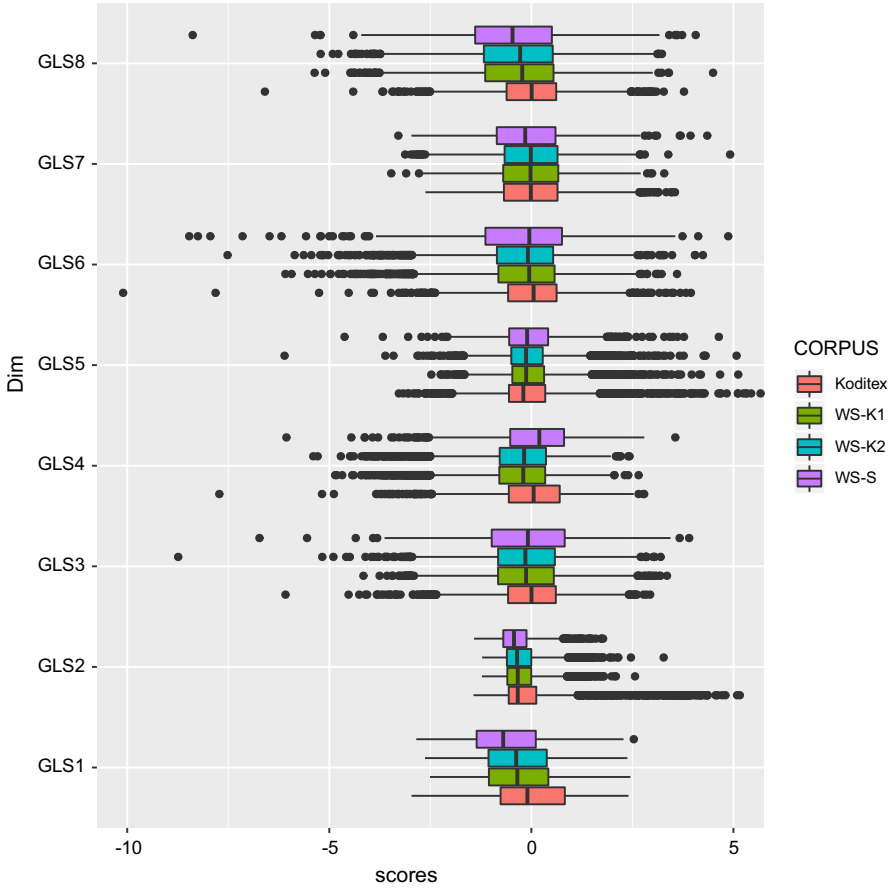

Fig. 2 Factor scores in all 8 dimensions for texts from Koditex (representing the original MD model) and Araneum samples WS-K1, WS-K2 and WS-S

Since we are interested in the ranges of variation covered by the corpora under examination, we should compare the span of sufficiently wide percentile ranges of scores observed across corpora in each dimension. While it is obvious that extreme outliers may be influenced by many factors unrelated to the topic of this study and can possibly bias the results, it is also clear, on the other hand, that comparing only the central parts of the distributions (such as interquartile ranges) would drastically curtail the inherent variability present in the data. In making this decision, it is worth considering that Koditex is a highly diverse corpus consisting of 3334 text chunks in 45 text classes (each containing cca 200,000 words in 67–91 chunks, with a median of 71); if interquartile ranges were used, entire classes could be excluded from the comparison as outliers. As a compromise, we decided to compare fairly wide (but not full) ranges of variation between the 2nd and 98th percentile.[9] Given the diversity of Koditex and relative homogeneity of the WS-* corpora, if any bias is introduced by this decision, it is in favor of the Araneum samples—since Koditex attempts to cover a wider range of different sources, the coverage is necessarily sparser, which means removing outliers has a more noticeable effect. The resulting intersections, Koditex complements and Araneum complements (see Fig. 1 for a visualization of these concepts) with respect to the joint range of variation covered by both types of corpora in each dimension are summarized in Table 4.

The interpretation of the data in Table 4 is quite straightforward. The proportion in the "Intersection" column indicates to what extent the Koditex corpus and Araneum sample are interchangeable within the particular dimension. The "X complement" columns contain the proportion of the variation range which is specific either to Koditex or one of the web samples (if the value is zero, the whole range of the corpus is covered by the other one, e.g. the range of variation of Koditex is wider in GLS5 than the range of any Araneum sample). To zero in onto what causes these differences, it might sometimes be helpful to compare the values of individual features within Koditex and the Araneum samples, respectively. We will specifically point out those cases where a feature salient for the dimension (i.e. one with a high loading) shows a statistically significant difference between Koditex and the WS-* corpora (at the 0.05 significance level), and the effect size of the difference is in the top decile. With this in mind, the following brief observations can be made about the individual dimensions:

5.1.1 Dimension 1: dynamic × static

Both types of corpora largely overlap in this dimension. Koditex brings in the positive (dynamic) extreme, whereas Araneum samples add texts widening the spectrum towards the negative pole (static). A significant difference can be identified in the 3rd person pronouns feature, which contributes to the dynamic perception of a text (related to narration) and which has a significantly higher

[9] The 2% limit is derived from the class quota in the Koditex corpus, which is at least 200,000 words, i.e. 2.21% of the corpus, or a minimum of 67 chunks, i.e. 2.04% of the corpus. This range ensures that even in the worst case scenario, none of the text classes in the Koditex corpus would be fully excluded as outliers.

Table 4 Relative proportions of the joint variation (i.e. Koditex + the Araneum sample in the given row) in each dimension covered by both types of corpora (intersection) vs. proportions covered by one type only (Koditex complement, Araneum complement)

Dimension	Araneum sample	Intersection with Koditex (%)	Koditex complement (%)	Araneum complement (%)
GLS1: dynamic × static	WS-K1	86.28	11.03	2.69
	WS-K2	84.34	11.41	4.25
	WS-S	75.84	13.01	11.14
GLS2: spontaneous × prepared	WS-K1	39.31	58.42	2.28
	WS-K2	39.26	58.28	2.46
	WS-S	40.48	53.95	5.57
GLS3: higher × lower level of cohesion	WS-K1	96.87	0.00	3.13
	WS-K2	94.44	0.00	5.56
	WS-S	73.65	0.00	26.35
GLS4: polythematic × monothematic	WS-K1	82.52	4.44	13.04
	WS-K2	82.86	5.38	11.76
	WS-S	86.52	0.00	13.48
GLS5: higher × lower amount of addressee coding	WS-K1	70.53	29.47	0.00
	WS-K2	69.96	30.04	0.00
	WS-S	88.47	11.53	0.00
GLS6: general × particular	WS-K1	79.48	1.50	19.02
	WS-K2	78.71	2.18	19.11
	WS-S	60.60	0.00	39.40
GLS7: prospective × retrospective	WS-K1	89.76	10.24	0.00
	WS-K2	88.61	11.39	0.00
	WS-S	95.33	0.06	4.62

Table 4 continued

Dimension	Araneum sample	Intersection with Koditex (%)	Koditex complement (%)	Araneum complement (%)
GLS8: attitudinal × factual	WS-K1	79.45	1.56	18.99
	WS-K2	78.33	2.28	19.39
	WS-S	66.64	0.00	33.36

Each corpus is represented by a 2–98 percentile range of factor scores in each dimension

median in Koditex than in the web samples (although overall outliers of this feature come from the WS corpora).

5.1.2 Dimension 2: spontaneous × prepared

Dimension 2 yields the smallest intersection between Koditex and the web samples, and consequently, the most salient difference between the two types of corpora. The Koditex complement introduces a range of texts on the positive (spontaneous) extreme from the spoken interactive category (*spo-int*). The most salient difference among features important for this dimension can be observed with respect to contact expressions, which signalize spontaneity. Contact expressions have both a higher median and also higher dispersion in Koditex, which amplifies the Koditex complement in this dimension.

5.1.3 Dimension 3: higher × lower level of cohesion

The large overlap in this dimension suggests that in general, web-crawled and traditional corpora do not differ substantially in the level of cohesion. As far as the complements are concerned, it is the only dimension where the range of variation covered by Araneum samples is greater than the original MD model based on Koditex. Texts gravitating towards both extremes—negative (less cohesive) as well as positive (more cohesive)—are introduced by Araneum samples. This is caused i.a. by predicative nouns and numerals: both features have a higher median in WS-* than in Koditex, which stretches the extremes of the WS corpora beyond the range covered by Koditex, because the former feature has a positive loading (causing texts to be more cohesive) and the latter a negative one (signaling lower cohesion).

5.1.4 Dimension 4: polythematic × monothematic

Another large overlap, this time complemented by some WS texts which extend the range of variation toward the negative values (monothematic). However, the most extreme texts (outliers, not taken into account in Table 4) are within the Koditex category *wri-nfc-adm* (administrative documents). Three features salient for this dimension have significantly different values in Koditex and in the web samples: Yule's coefficient (representing lexical repetitiveness), passive voice and verbal nouns. All of them have negative loadings, which means the higher they are, the more monothematic a text is. In all three cases, WS texts tend to have a higher median value. As far as the extreme values are concerned however, only one out of the three (passive voice) has considerably higher outliers in the WS-* corpora; as for the remaining two features, WS-* and Koditex outliers do not differ significantly.

5.1.5 Dimension 5: higher × lower amount of addressee coding

The overlap between corpora in this dimension is somewhat smaller. It is one of the dimensions where Araneum samples do not widen the range of variation (their

variation is fully covered and even surpassed by Koditex texts). Koditex contains texts with higher amounts of addressee coding (i.e. the positive pole), mainly screenplays and dialogues in fiction. The differences between salient features in this dimension are not as pronounced (none of the features mentioned below exhibits a difference with an effect size within the highest decile). Nevertheless, the difference between Koditex and WS texts in this dimension is driven by features related to questions (non-polar questions and questions in general) and average sentence length (SL). While the question-related features have positive loadings and thus contribute to a higher amount of addressee coding, the effect of SL is quite the opposite. It is therefore not surprising that the median frequency of questions is significantly higher in Koditex, whereas median SL is lower. However, in all three cases, the distribution is highly skewed with distant outliers in some WS texts; especially the SL feature is questionable here, because one of the outliers is an obvious case of erroneous automatic segmentation (which is to be expected in web-crawled texts).

5.1.6 Dimension 6: general × particular

A moderate overlap between corpora in this dimension is accompanied by a WS complement which introduces texts widening the coverage towards the negative extreme (focus on particular referents). The only highly salient feature for this dimension which simultaneously exhibits a statistically significant difference between corpora are numerals, which contributes to the particular and specific pole of this dimension and which has already been discussed above under dimension 3.

5.1.7 Dimension 7: prospective × retrospective

Another large overlap. Apart from the short Araneum samples, no new variation is added by the WS texts to the original MD model; Koditex complements, on the other hand, can be found on both extremes (prospective as well as retrospective). The only relevant feature for comparison in this dimension is 3rd person pronouns, which contributes to the retrospective extreme and has already been discussed in dimension 1.

5.1.8 Dimension 8: attitudinal × factual

A moderate overlap in this dimension is accompanied by a WS complement. According to the MD model, the web samples in general tend to score in the lower regions of this dimension, leaning towards the factual extreme. There are no significant differences in salient features to report for this dimension.

5.2 Text length and dispersion

Before proceeding to the 2-dimensional comparison, the data from the Araneum corpus provide a unique opportunity to examine the issue of text length in MDA and by extension also in corpus design in general. As suggested by the results presented above, specifically the contrast between WS-S and WS-K1/2 apparent from Table 4,

Table 5 Average proportions of shared and corpus-specific variation ranges

Araneum sample	Intersection with Koditex (%)	Koditex complement (%)	Araneum complement (%)
WS-K1	78.00	14.60	7.39
WS-K2	77.10	15.10	7.82
WS-S	73.40	9.82	16.70

the shorter the texts in a data set, the greater the dispersion of their factor scores. First, let us look at the average joint and corpus-specific variation for all dimensions as summarized in Table 5.

While Table 4 reveals that the Araneum sample with ten times shorter texts (WS-S) has consistently higher variation than the other two Araneum samples (WS-K1 and WS-K2), as shown by the figures in the "Araneum complement" column for each triplet of rows corresponding to a given dimension, comparing the average numbers in the "X complement" columns of Table 5 suggests that WS-S adds more to the joint variation than even the Koditex corpus (16.7% vs. 9.82%). Conversely, the other batches WS-K1 and WS-K2 (with text length distributions modeled according to Koditex) add less. Since all three WS corpora are comparable in all relevant aspects except for text length distribution (they were all obtained by random sampling of the same type of source data, just sliced into longer or shorter chunks), this hints at a relationship between text length and dispersion of factor scores, which is our proxy for variation coverage.

This relationship can be explained by the following rationale: shorter texts provide less surface area for the co-occurrence of features, and thus they are inevitably more homogenous register-wise. When projected onto the MD model, this translates into an inclination towards dimension extremes. It is obvious that e.g. 10 occurrences of 2nd person pronouns scattered over 20,000 words in a novel contribute less to a perception of addressee coding, than the same amount of these pronouns in an excerpt of 2000 words. Thus, in a shorter text, each occurrence of a feature has more weight in profiling the text and, at the same time, shorter texts provide less opportunity to level this out by containing other features with the opposite effect. Mathematically speaking, shorter texts tend to yield more extreme relative frequencies (each attested occurrence counts a lot, and conversely, many features will be completely absent), which then translates into more extreme factor scores (= extreme positions in the MD space) and more pronounced characteristics register-wise. On the other hand, longer texts covering a wider array of features with moderate relative frequencies tend to gravitate towards the dimension centers.

This account is empirically supported by comparing the Araneum samples with Bartlett and Fligner-Killeen tests of homogeneity of variances (see Table 6). Since the tests are sensitive to the position of the median, all the data were re-centered around the same median.

In all dimensions, the 2–98 percentile variation ranges covered by WS-S are wider than those covered by WS-K1 and WS-K2 (regardless of whether the difference was significant). The exact proportions are listed in the "Dispersion rate"

Table 6 *p*-values of Bartlett and Fligner-Killeen tests of homogeneity of variances comparing the Araneum batches of longer texts (WS-K1 and WS-K2) with the one with shorter texts (WS-S)

Dimension	WS-K1 vs. WS-S			WS-K2 vs. WS-S		
	Bartlett	Fligner-Killeen	Dispersion rate WS-K1/WS-S	Bartlett	Fligner-Killeen	Dispersion rate WS-K2/WS-S
GLS1: dynamic × static	0.1425	0.4445	0.9339	0.1157	0.3391	0.9452
GLS2: spontaneous × prepared	0.0154	0.5079	0.8726	0.0183	0.415	0.8771
GLS3: higher × lower level of cohesion	< 0.0001	< 0.0001	0.7603	< 0.0001	< 0.0001	0.7798
GLS4: polythematic × monothematic	< 0.0001	< 0.0001	0.9507	< 0.0001	< 0.0001	0.9278
GLS5: higher × lower amount of addressee coding	< 0.0001	< 0.0001	0.7972	< 0.0001	< 0.0001	0.7908
GLS6: general × particular	< 0.0001	< 0.0001	0.7371	< 0.0001	< 0.0001	0.7328
GLS7: prospective × retrospective	< 0.0001	< 0.0001	0.8566	< 0.0001	< 0.0001	0.8456
GLS8: attitudinal × factual	< 0.0001	< 0.0001	0.8098	< 0.0001	< 0.0001	0.8079

"Dispersion rate" is calculated as the proportion of the 2–98 percentile range of WS-S covered by the 2–98 percentile range of WS-K1 or WS-K2, and serves as a trivial effect size estimator. When close to 1, the dispersions are similar; the smaller it is, the wider the dispersion of WS-S compared to WS-K1 or WS-K2

columns in Table 6. Even though the differences in dispersion may not be that large (the dispersion rate in Table 6 reveals that the dispersion of WS-K1 and WS-K2 is between 73–95% of the dispersion of WS-S, depending on the dimension), these results show that text length plays an important role in the interpretation of MD models. MD models based on corpora with an uneven distribution of text lengths may lead to skewed positions of those groups of texts which are customarily shorter and therefore tend to occupy more extreme positions (e.g. letters or administrative documents), compared to those that are longer and gravitate towards dimension centers (e.g. novels; cf. Cvrček et al. 2018b, Sect. 2.1).

On the other hand, the results in Table 6 also show that the difference in dispersion is significant in all but the two most important dimensions explaining the largest amount of variance in the model (GLS1 and GLS2). This leads to two general remarks about MDA: first, text length as a factor influencing the dispersion of factor scores is important especially within dimensions focusing on more specific textual features (rather than very general dimensions like GLS1 and GLS2). Second, the two most important dimensions are also the most stable ones (with respect to the dispersion of texts). This is likely due to the fact that the key dimensions of an MD model (those that explain the most variance) are based on frequently occurring features (or a number of features structurally closely related) whose estimates are robust and stable across long stretches of text.

Even more generally, the fact that shorter texts are prone to deviations in frequencies of features, which then leads to greater differences between texts in MD space, also has an important consequence for corpus design in general. Following the findings described above, we may infer that corpora with shorter texts (typically text excerpts) can be viewed as instantiating distinctive registers better than a corpus of the same size with longer texts. It may be an important aspect to consider especially in situations in which a smaller corpus is required, yet it is important to cover as much language variation as possible. By using excerpts, such a corpus may adequately represent a wide range of highly pronounced registers, while, on the other hand, a possible disadvantage is that it might fail in representing texts of mixed registers (which are expected to exist in real life communication).

5.3 2-dimensional comparison

A comparison of the variability covered by Koditex and the Araneum samples on the basis of the full MD model can be difficult to grasp intuitively, as it involves no less than 8 dimensions of variation. Moreover, the amount of variance explained by the individual dimensions is not equal, as can be seen in Table 2, so not all dimensions are equally relevant. Considering the findings from the previous section—that the two most important dimensions are also the most stable ones—it seems justifiable to perform a simplified but more intuitive comparison limited to the first two dimensions only.

As a brief refresher on the types of variation accounted for by the first two dimensions, we turn to characteristics given in previous research (but cf. also Sect. 2.2). Dimension 1 ("GLS1" in the figures, dynamic vs. static), represents the "verbal vs. nominal distinction [which] is associated predominantly with written

genres (especially romance novels and letters on the positive side). With past tense verbs as the most salient verbal feature, we can conclude that our dimension 1 marks both personal involvement and narration in contrast to description" (Cvrček et al. 2018b, Sect. 5.9). The second dimension ("GLS2", spontaneous vs. prepared) "isolates spoken, interactive and private genres. The combination of features marking (1) interactivity and online production (contact expressions, fillers, demonstratives, word repetition) and (2) informality (expressive particles, interjections) attract (3) conventionalized non-standard Common Czech morphonological variants, symptomatic of diglossia." (Cvrček et al. 2018b, Sect. 5.9).

The visual 2D comparison in Figs. 3 and 4 shows the positions of all the texts from Koditex and all Araneum samples along the first two dimensions. Each text is represented by a shaded dot, and colored ellipses mark areas occupied by selected groups of interest.

The area occupied by the web samples (indicated by dashed ellipses) lies in the lower part of the figures, where non-fiction and journalism can be found in the original MD model. In Fig. 3, three full ellipses mark non-web text classes from Koditex which occupy noticeably different regions of the chart (the ellipses represent 95% confidence intervals, therefore some outliers may be outside the area). Most conspicuous among these is the category of informal spoken dialogues (*spo-int*), which does not overlap with the Araneum area at all. The remaining two—fiction (*wri-fic*) and private correspondence (*wri-pri*)—exhibit some overlap but different centers of gravity, and in addition, they introduce texts clearly beyond the range of Araneum samples.

All three of these Koditex-specific regions are dynamic (they contain above-average rates of verbal features such as the indicative and the past tense, or verb-related features such as 3rd person pronouns). These features are used primarily for narration, contemplation (another salient feature is verbs of thinking), or for the clause chaining structure typical of spontaneous texts. Conventionally, the first use would be associated with fiction and the latter two with private correspondence and conversation, but in practice, the boundaries can be blurry. As for the spontaneous vs. prepared dimension, two of the three Koditex-specific regions tend towards the spontaneous extreme (top): most notably, unprepared spoken conversation, which has the highest scores on this dimension among all text categories in this study. But even fiction texts, which seem to have a moderate position with respect to the 2nd dimension, are more spontaneous than web sample texts, presumably due to mimicking spontaneous speech in reported speech, or as part of the author's style.

In short, for these specific text categories, no satisfactory substitute can be found on the searchable web. Therefore, any research on Czech using an opportunistic web-crawled corpus should be aware of the fact that the data do not cover a functionally important part of the variability.

It is probably no coincidence that the areas of the chart conspicuously not covered by the web samples correspond to those types of texts which are difficult to obtain. In contrast to mainly journalistic texts, which are available in abundance on the web and also elsewhere in the offline world (e.g. news-monitoring agencies), text categories such as prototypical spoken discourse (i.e. unprepared, spontaneous,

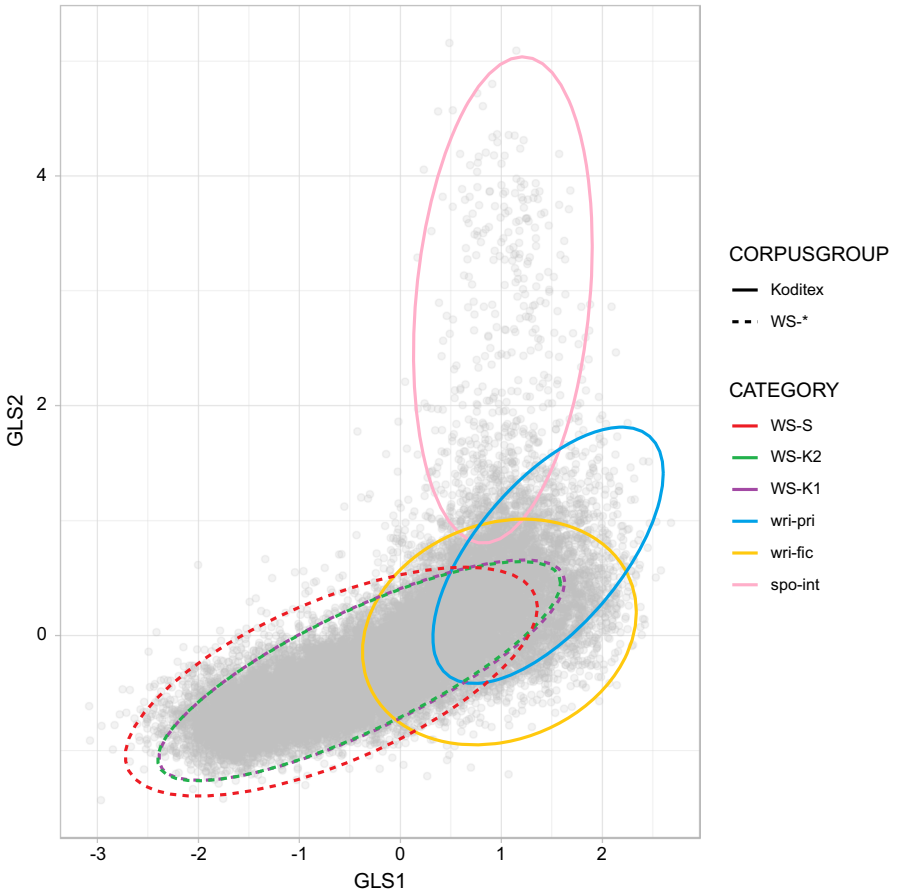

Fig. 3 Koditex classes (full contours) which occupy noticeably different regions of the space defined by the first 2 dimensions of the MD model compared to web-crawled Araneum samples (WS-*, dashed contours)

informal etc.), private correspondence, or some types of fiction are hard to get in the real world as well as on the web.

Finally, we examine the *web* mode Koditex categories separately, so as to emphasize a distinction between web language in general and web-*crawlable* language. As mentioned previously, general web-crawling techniques can only access the "searchable" web; any type of content which requires authentication or a custom API to access is out of bounds. Unfortunately, this excludes a lot of user-generated content on social networks and similar platforms: for one thing, access to this data is often restricted in the ways described above, and even if it is not, these pages can be noisy, with short bursts of content interspersed with automatically generated metadata, so that they risk being discarded by automatic filters, which are necessary in web crawling to get rid of garbage data. This is unfortunate, as these platforms are extremely linguistically interesting, with vast amounts of participating

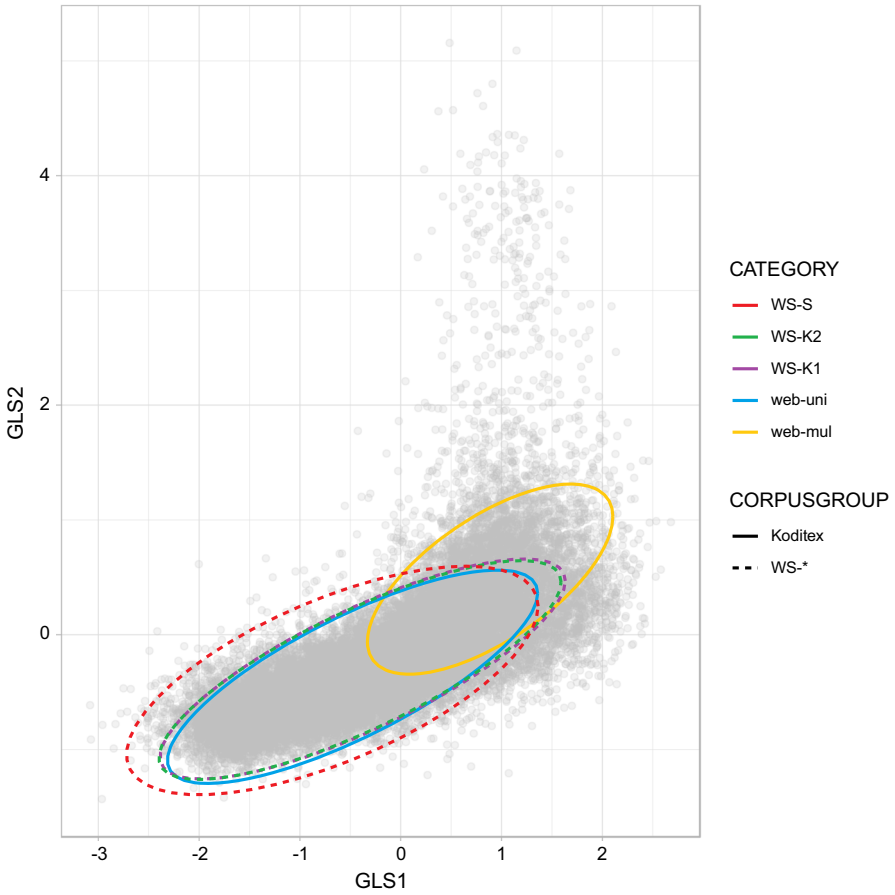

Fig. 4 Comparison of web-based Koditex text classes (full contours) with Araneum web-crawled data (WS-*, dashed contours) in 2D space

speakers/authors. By contrast, when designing Koditex as a manually curated corpus with a focus on diversity, we sought alternative, tailored sources for this type of content which is usually out of the reach of crawling engines (cf. Sect. 2.1). What we obtained is far from being exhaustive, but it provides us with at least some comparison.

The regions of web content covered by Koditex and the WS corpora are compared in Fig. 4, which suggests that the Koditex *web-uni* class (unidirectional communication on the web—blogs, Wikipedia articles) is fully covered by the opportunistic approach adopted in the Araneum corpus. After all, Araneum is the source of the blog samples included in Koditex, so this is not that surprising. On the other hand, the *web-mul* class (multidirectional web-based communication—comments sections, Facebook posts, forums) is partially beyond the coverage of the Araneum batches and, more importantly, it represents a distinct area of variability.

As a consequence, there might even be specific web genres (with unique register characteristics) which are eluding web harvesting and are better represented in specialized corpora (e.g. a corpus of Twitter posts) than in general opportunistic web-crawled corpora.

All of this being said, the overlap of other Koditex classes with Aranea samples is usually greater (see above in Table 4), therefore we can conclude that as far as the first two dimensions are concerned, the remaining classes can be more or less satisfactorily represented by web-crawled texts. E.g. text categories like public speeches (*spo-nin*), newspapers and magazines (*wri-nmg*) or non-fiction (*wri-nfc*, for details see Table 1) are fully covered by the texts from Araneum samples. With respect to the two most important dimensions of the MD model, these text categories can be fully substituted by web-crawled content without reducing the amount of variation in the data. It should however be pointed out that when this is done, other caveats to web-crawled data still apply (namely paucity of reliable metadata).

6 Conclusion

This study presents a comparison of two types of corpora on the basis of an MD model of register variation: Koditex, a carefully designed "traditional" corpus of Czech texts ranging from spoken to web communication to the written mode, assembled with an emphasis on as diverse a composition as possible, versus several samples of an opportunistic web-crawled corpus representing the Czech "searchable" internet. Projecting the web data onto the MD model, which is based on the former corpus, allows us to quantify the overlap between the corpora and identify their specificities. By comparing the range of variation covered by individual (sub)corpora on each dimension, we can estimate to what extent the traditional corpus is replaceable by the web-crawled corpus, which is of course cheaper and easier to obtain.

Crucially, this comparison is not based on extratextual metadata, but on the intratextual features which form the backbone of the MD model. Thus, when we conclude e.g. that text category X is not covered by web-crawled data, we do not simply mean that there are no web texts labeled in their metadata as X. Indeed, should X happen to be a non-web category by definition, then the conclusion that it is not available in web-crawled data would be trivial and not worth reporting. No; what we mean instead is that there is not even an acceptable substitute for X in web-crawled data, i.e. texts which would draw on similar linguistic resources, perform similar functions and bear witness to the same niche of language use as category X, irrespective of any external metadata.

The results show that the overlap is generally large in those areas of linguistic variation which correspond to text categories which are easy to obtain on the web as well as when building an offline-text corpus (namely journalistic texts and non-fiction texts). Although web-crawled texts do occasionally tend towards their own distinctive regions of the MD space (texts which are static, less cohesive, factual and

focused on particular referents),[10] unique text categories occupying distinct areas are only found in the traditional Koditex corpus. These are namely spoken informal (intimate) discourse, written private correspondence and some types of fiction (dynamic and addressee oriented).[11] The case of multi-directional classes of web communication (forums, Facebook posts etc.) then documents that while some areas of variation may actually be covered by web *content*, they may still be outside the reach of generalistic web-*crawlers*. In summary, all of these text categories cannot be substituted by general web-crawled data. It is thus important to build and maintain resources of this kind in order to secure full coverage of linguistic variation in empirical research.

According to these findings, the belief with which some researchers approach web-crawled corpora—that they can be considered as (nearly) faithful representations of the entirety of a given language simply by virtue of their unprecedented size—is inappropriate. To come back to the example used in the introduction—is it reasonable to use a web-crawled corpus as a representation of both oral and literate discourse? Based on the results presented in this paper, we can dismiss such an assumption as unwarranted, as none of the web-crawled texts show linguistic characteristics that would mimic those of actual spoken interactions. Any research based solely on web-crawled data is thus at risk of excluding important portions of variation patterns which can be found exclusively (in the case of Czech) in spontaneous interaction (oral, or written both on- and offline), and in some kinds of fiction. It goes without saying that the amount of risk varies depending on the topic, method and scope of the research question, e.g. studies pertaining to variation may be more sensitive to this issue, while many other research topics may be influenced only partially.

Despite being somewhat tangential to the results presented above, a crucial question regarding the generalizability of the present research is whether we can extend any of the conclusions from this study to other languages, especially English. It is obvious that the English internet, with its unparalleled population of active users (both L1 and L2 speakers) and consequently enormous textual online production, is in a unique situation, an outlier among other languages.[12] We can therefore assume that any other language used on the web is more directly comparable with Czech than English. However, our findings are, in general, in line with previous research in this domain done on English. Ide et al. state that as far as the English internet of the year 2002 is concerned, there is an abundance of "dense type of prose characteristic of formal documents", while there is lack of the "face-to-face conversation or other spoken interactions" (2002, p. 842). As mentioned previously, this could have been due to their restricted crawling which focused only

[10] Notice that these properties are characteristic of a catalog (list, table, or other condensed data presentation format) and similar text categories. These are indeed abundant on the web. They are also not very linguistically interesting.

[11] Again, the claim is *not* that there is e.g. no *actual* written private correspondence in web-crawled data (that would hardly be surprising), but that the web-crawled data does not even yield texts which would be linguistically equivalent to and could act as surrogates for private correspondence.

[12] See e.g. https://en.wikipedia.org/wiki/Languages_used_on_the_Internet (visited October 2019).

on institutional *.gov* and *.edu* domains, but given the similar results obtained on our more inclusive web-crawled data, we find it unlikely.

As a byproduct, the analysis helped in testing a hypothesized relationship between the text length distribution in a corpus and the dispersion of the texts' factor scores in an MD model. Shorter texts seem to be less heterogenous and more pronounced in their characteristics than longer ones, therefore they gravitate more towards the extremes of dimensions. As a consequence, corpora used for MDA should contain texts (or preferably, text excerpts) of comparable lengths in order to mitigate this influence.

Acknowledgements This study was supported by the European Regional Development Fund project "Language Variation in the CNC" no. CZ.02.1.01/0.0/0.0/16_013/0001758 and has been, in part, funded by the Slovak KEGA and VEGA Grant Agencies, Project No. K-16-022-00 and 2/0017/17, respectively.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Anthony, L. (2018). *AntCorGen*. Tokyo: Waseda University. Retrieved November 23, 2018, from http://www.laurenceanthony.net/software.

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British national corpus with SARA*. Edinburgh: Edinburgh University Press.

Baron, N. (2010). *Always on: Language in an online and mobile world* (1st ed.). Oxford: Oxford University Press.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation, 43*(3), 209–226. https://doi.org/10.1007/s10579-009-9081-4.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. In *Proceeding of the EuraLex Conference* (pp. 123–132).

Benešová, L., Křen, M., & Waclawicová, M. (2013). *ORAL2013: Representative corpus of informal spoken Czech*. czech, Praha: Institute of the Czech National Corpus. FF UK. Retrieved March 18, 2020, from http://www.korpus.cz.

Benko, V. (2014). Aranea: Yet another family of (comparable) web corpora. In *International Conference on Text, Speech, and Dialogue* (pp. 257–264). Berlin: Springer.

Benko, V. (2016a). Two years of Aranea: Increasing counts and tuning the pipeline. In *LREC* (pp. 4245–4248).

Benko, V. (2016b). Feeding the "Brno Pipeline": The case of Araneum Slovacum. *RASLAN 2016 Recent Advances in Slavonic Natural Language Processing, 10,* 19–27.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–257.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast, 14*(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib.

Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics, 44*(2), 95–137. https://doi.org/10.1177/0075424216628955.

Čermák, F., Adamovičová, A., & Pešička, J. (2001). *PMK: Prague spoken corpus*. czech, Praha: Institute of the Czech National Corpus. FF UK. Retrieved March 18, 2020, from http://www.korpus.cz.

Cvrček, V., Čermáková, A., & Křen, M. (2016). Nová koncepce synchronních korpusů psané češtiny. *Slovo a slovesnost, 77*(2), 83–101.
Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018a). Variabilita češtiny: multidimenzionální analýza [Variability of Czech: A multi-dimensional analysis]. *Slovo a slovesnost, 79*(4), 293–321.
Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018b). From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2018-0020.
Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (forthcoming). *Register variability of elicited texts*.
Davies, M. (2018). *The 14 Billion Word iWeb Corpus*. Retrieved May 10, 2019, from https://www.english-corpora.org/iweb/.
Francis, W. N., & Kučera, H. (1964, 1979). Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. *Brown Corpus Manual*. Retrieved December 13, 2018, from http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM.
Górski, R. L., & Łaziński, M. (2012). Reprezentatywność i zrównoważenie korpusu. In A. Przepiórkowski, M. Bańko, R. L. Górski, & B. Lewandowska-Tomaszczyk (Eds.), *Narodowy korpus języka polskiego: praca zbiorowa* (pp. 25–36). Warszawa: Wydawnictwo Naukowe PWN.
Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430–450.
Herring, S. C. (2010). Computer-mediated conversation Part I: Introduction and overview. *Language@internet*, *7*(2). Retrieved March 18, 2020, from https://www.languageatinternet.org/articles/2010/2801.
Hladká, Z. (2002). *BMK: Brno spoken corpus*. Praha: Institute of the Czech National Corpus. FF UK. Retrieved March 18, 2020, from http://www.korpus.cz.
Hoffmannová, J., Homoláč, J., Chvalovská, E., Jílková, L., Kaderka, P., Mareš, P., et al. (2016). *Stylistika mluvené a psané češtiny* (1st ed.). Praha: Academia.
Ide, N., Reppen, R., & Suderman, K. (2002). The American National Corpus: More Than the Web Can Provide. In *Proceedings of the Third Language Resources and Evaluation Conference (LREC)* (pp. 839–844). Presented at the LREC 2002, Las Palmas, Canary Islands, Spain: Citeseer. Retrieved March 18, 2020, from http://www.lrec-conf.org/proceedings/lrec2002/pdf/303.pdf.
Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The tenten corpus family. In *7th International Corpus Linguistics Conference CL* (pp. 125–127).
Kaderka, P. (2012). *Dialog: corpus of broadcasted Czech discussions*. czech, Praha: Ústav pro jazyk český, AV ČR. Retrieved March 18, 2020, from http://www.korpus.cz.
Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics, 6*(1), 97–133.
Kilgarriff, A. (2012). Getting to know your corpus. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, speech and dialogue* (pp. 3–15). Berlin: Springer.
Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P. V. S. (2010). A corpus factory for many languages. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17–23 May 2010, Valletta, Malta* (pp. 17–23). Valleta, Malta. Retrieved March 18, 2020, from http://www.lrec-conf.org/proceedings/lrec2010/summaries/79.html.
Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Jelínek, T., et al. (2016). SYN2015: Representative corpus of contemporary written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 2522–2528). Presented at the LREC'16, Portorož: ELRA.
Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 133–149). Amsterdam: Rodopi.
Michelfeit, J., Pomikálek, J., & Suchomel, V. (2014). Text tokenisation using unitok. In *8th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU* (pp. 71–75). Presented at the RASLAN 2014, Brno: NLP Consulting.
Piperski, A. (2017). *Sum of Minimum Frequencies as a Measure of Corpus Similarity*. Presented at the Corpus Linguistics 2017, Birmingham. Retrieved March 18, 2020, from https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper143.pdf.
Piperski, A. (2018). Corpus size and the robustness of measures of corpus distance. In *Computational Linguistics and Intellectual Technologies* (pp. 590–600). Presented at the Dialogue 2018, Moscow. http://www.dialog-21.ru/media/4327/piperskiach.pdf.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora* (PhD Thesis). Masarykova univerzita, Fakulta informatiky, Brno. Retrieved March 18, 2020, from https://is.muni.cz/th/o6om2/phdthesis.pdf.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved March 18, 2020, from https://www.R-project.org/.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora—Volume 9* (pp. 1–6). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1117729.1117730.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, IL: Northwestern University. Retrieved March 18, 2020, from https://CRAN.R-project.org/package=psych.

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora, 13*(1), 65–95. https://doi.org/10.3366/cor.2018.0136.

Suchomel, V., & Pomikálek, J. (2012). Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)* (pp. 39–43). Lyon.

Válková, L., Waclawicová, M., & Křen, M. (2012). Balanced data repository of spontaneous spoken Czech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3345–3349). Presented at the LREC'12, Istanbul: ELRA. Retrieved March 18, 2020, from http://www.lrec-conf.org/proceedings/lrec2012/pdf/179_Paper.pdf.

Zasina, A. J., & Komrsková, Z. (2019). Koditex — korpus diverzifikovaných textů. *Studie z aplikované lingvistiky - Studies in Applied Linguistics, 10*(1), 127–132.

Zasina, A. J., Lukeš, D., Komrsková, Z., Poukarová, P., & Řehořková, A. (2018). *Koditex: corpus of diversified texts*. Czech, Prague: Institute of the Czech National Corpus. FF UK. Retrieved November 26, 2018, from http://www.korpus.cz.