Václav Cvrček*, Zuzana Komrsková, David Lukeš,
Petra Poukarová, Anna Řehořková and Adrian Jan Zasina

From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA

https://doi.org/10.1515/cllt-2018-0020

Abstract: This paper is part of a larger research effort on language variability aimed at uncovering the relations between extra- and intratextual characteristics of Czech texts by means of multi-dimensional analysis. The palpable lack of prior art on quantitative register analysis of Czech led to several distinctive methodological decisions, concerning namely corpus design, feature selection and the parameters of factor analysis, especially the number of dimensions to extract. We report on these for their potential relevance to other researchers embarking on a similar journey. In order to demonstrate the viability of the model, we also present a brief interpretation of the resulting dimensions.

Keywords: multi-dimensional analysis, register variation, methodology, Czech

1 Introduction

As one of the most widely used methods in describing language variation, multi-dimensional analysis (MDA) does not need a particularly detailed introduction. Since its inception at the hands of Douglas Biber (1988), MDA has shed light on

***Corresponding author: Václav Cvrček,** Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic, E-mail: vaclav.cvrcek@ff.cuni.cz
http://orcid.org/0000-0003-3977-2393
Zuzana Komrsková: E-mail: zuzana.komrskova@ff.cuni.cz, **David Lukeš:** E-mail: david.lukes@ff.cuni.cz, **Petra Poukarová:** E-mail: petra.poukarova@ff.cuni.cz, **Anna Řehořková:** E-mail: anna.rehorkova@ff.cuni.cz, **Adrian Jan Zasina:** E-mail: adrian.zasina@ff.cuni.cz, Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic
http://orcid.org/0000-0002-1170-9344
http://orcid.org/0000-0003-0429-6542
http://orcid.org/0000-0003-3707-6466
http://orcid.org/0000-0002-6676-317X
http://orcid.org/0000-0001-9348-5833

Published online October 20, 2018

variation in a typologically diverse handful of languages, chief among those English, which has repeatedly been under scrutiny by Biber himself and others (Biber 1990, 1995; Biber and Conrad 2009; Biber and Egbert 2016), in broad and narrow focus. This paper introduces an MDA of Czech, a West Slavic language with a sociolinguistic situation bordering on diglossia (see e.g. Bermel 2014 for a recent overview) which has never been fully explored in this way.

As MDA matured as a framework, a natural feedback loop emerged whereby the design of more recent multi-dimensional (MD) studies is informed by the results of earlier ones (which for obvious reasons cannot be the case of Czech yet). Case in point: corpus design which, in MDA, is always guided by a concern for representative sampling of the registers the analysis aims to cover. However, the meaning of the term "register" itself has developed over time. In an early book, "register distinctions are defined in non-linguistic [= extratextual] terms" (Biber 1995: 7), but as regular patterns of co-occurrence of extra- and intratextual qualities in a given language are uncovered through successive MDAs, they become a new starting point for further ventures, so much so that by the time of Biber and Conrad (2009: 2), "the register perspective combines an analysis of *linguistic characteristics* that are common in a text variety with analysis of the *situation of use* of the variety" (emphasis added).

The authors go on to state explicitly that "the process of register analysis is often iterative" (Biber and Conrad 2009: 10), as broad extratextual classifications are gradually refined by intratextual evidence. Clearly, this holds true not only within a single study, but particularly across multiple studies, spanning potentially decades.[1]

At the outset of our MDA of Czech, we found ourselves very much at the start of this iterative process. Apart from an unpublished proof-of-concept study by Vilém Kodýtek, which used a limited amount of features based on Biber's (1988) model and applied them to a less diverse corpus, there was a dearth of quantitative research into Czech register variation. Bringing MDA to a new language calls at the very least for language-specific features which have to be devised from scratch (with the help of extant language descriptions) and which represent a methodological innovation *sui generis*. At the same time though, we were fortunate enough to be able to draw upon methodological insights from many MDAs of other languages.[2]

1 Cf. also recent work by Biber and Egbert (2016), who defined registers in terms of narrativity, informativity, etc., i.e. according to linguistic and functional characteristics of the texts.

2 We are also indebted to Douglas Biber and Václav Březina, who kindly agreed to devote a substantial amount of time to reviewing our progress at crucial junctures.

As a result, our guiding principle for this first iteration of Czech MDA was to chart the space of variation, bootstrapping from extratextual characteristics into the intratextual (cf. also Lee 2001: 62). The paper documents where we strayed from established practice and argues why such choices might make sense, especially when bringing MDA to a new language with little prior research to act as a safety net against systematic error. We begin by discussing the design of our corpus and related issues. Then, we give an account of the linguistic features employed, with a particular emphasis on type-based features. A separate section is devoted to the statistical procedure at the heart of MDA, factor analysis (FA), focusing on ways to establish the number of dimensions to interpret. Since the proof of the pudding is in the eating, we end with a quick overview of the resulting dimensions of variation of Czech (for a more detailed account, incl. the full list of features used, cf. Cvrček et al. Forthcoming). We stop just shy of synthesizing extra- and intratextual characteristics into a description of registers in Czech and leave that to future work.

2 Corpus design

A corpus for MDA consists of entities whose linguistic characteristics are measured and subsequently compared to each other to yield underlying dimensions of variation. Traditionally, these entities have been mapped directly onto the concept of *text*, but for reasons that will become clear below, we will refer to them using the more general statistical term *observation unit* for now. Two key questions come to mind when designing an MDA corpus:

1. What constitutes observation units, how should they be delimited?
2. What are the criteria for sampling the population of available units, i.e. deciding what to put in the corpus?

If the task at hand is to explore the previously uncharted space of register variation in a language, then a reasonable requirement is that the set of observation units should span the largest possible portion of this space. It is fairly obvious that this constrains point 2 above: the selected units should be linguistically varied, resulting in a heterogeneous corpus mapping the whole space. What is perhaps less obvious is that this requirement offers guidance with respect to point 1 above as well: the units themselves should be as linguistically homogeneous as possible. Both points are elaborated on below.

2.1 Homogeneity of observation units

A homogeneous unit can be defined as one that predominantly exhibits a restricted number of linguistic features associated with a coherent set of discourse or communicative functions. A heterogeneous unit, on the contrary, contains many features pertaining to various different functions and occurring at similar rates. Homogeneity is desirable in that such units will naturally tend to stretch out the extremes of the resulting dimensions, making them stand out more clearly from one another, which makes them easier to interpret. By contrast, mixed units will tend to huddle in the vicinity of the centroid of the multi-dimensional space, as the different influences they encompass cancel out.

It is a matter of statistical necessity that longer units are more likely to be mixed, purely by dint of offering more surface area for the occurrence of any feature and therefore also for feature co-occurrence. The first and foremost criterion for ensuring the homogeneity of observation units is therefore length. Ideally, the length should be such that it allows a reliably stable estimate of the values of linguistic features without introducing too much mixing. This means that observation units which are intuitively conceived of as texts will not necessarily do: a novel is much too long to avoid mixing, a tweet much too short for reliable estimation.[3] In practice, more tangible and actionable constraints will likely emerge from the nature of the available textual material and the stated goals of the analysis, but the pitfalls associated with both extremes should be kept in mind.

In any case, observation unit lengths should be in the same ballpark, to ensure a level playing field in terms of the probabilities of co-occurrence of features. If parts of the corpus were noticeably biased in favor of longer units, e.g. parts derived from fiction as opposed to private letters, they would also be systematically biased in favor of mixing, so dimensions primarily associated with fiction would be misrepresented (e.g. as having a smaller dispersion within the variation space). We ended up compromising on a length in the low thousands of words, 2,000–5,000 to be exact, with some exceptions going as low as 1,000 where data was scarce, but other ranges can make sense depending on the situation.

How to derive observation units from texts longer than that is easy: just chop them up into smaller segments (honoring sentence boundaries), call these *chunks*. As far as establishing dimensions (and consequently identifying registers) via MDA is

3 This is not to deny that novels have a register. The argument applies only to the process of staking out the space of variation.

concerned, working on chunks instead of whole texts presents no theoretical obstacle:

> Registers can be identified and described based on analysis of either complete texts or a collection of *text excerpts*. This is because the linguistic component of a register analysis requires identification of the *pervasive linguistic features in the variety*: linguistic characteristics that might occur in any variety but are much more common in the target register. (emphasis added; Biber and Conrad 2009: 6)

What about extremely short texts? In order to include them in the same analysis, there is only one option: *aggregation* according to some criteria designed to provide a different type of linguistic homogeneity, in place of the homogeneity which comes with being part of the same stretch of communication. One such criterion is authorship, which we exercised on Facebook and forum posts. Newspaper articles, on the other hand, where authorship attribution is often unavailable, were grouped by the section they appeared in (politics, sports, culture ...) within the respective periodicals. If the resulting aggregates are unnecessarily large, secondary criteria can be applied to achieve more fine-grained grouping; in our case, we used various kinds of temporal proximity between the original texts. A visual overview of the different procedures involved in producing our Koditex[4] corpus (Zasina et al. 2018) starting from source texts is given in Figure 1.

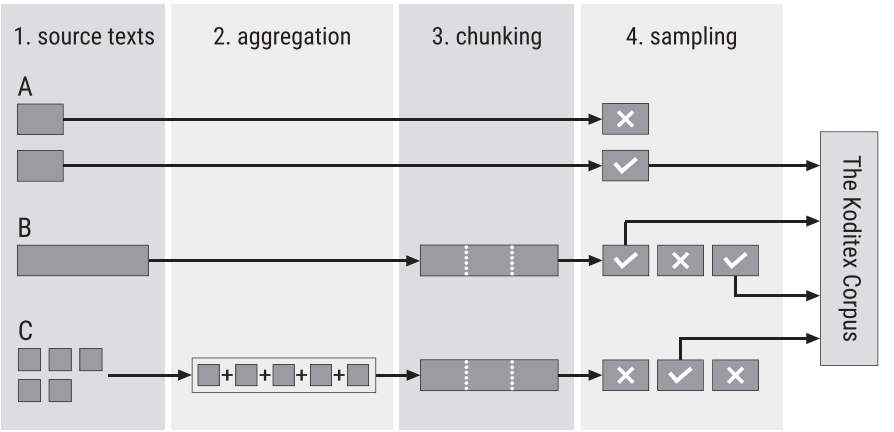

Figure 1: Building an MDA corpus consisting of text chunks of similar length. Stage 4 (sampling) is addressed in Section 2.2.

4 The name is both an acronym of the Czech version of the phrase *co*rpus of *di*versified *tex*ts and a tribute to Vilém Kodýtek.

An interesting category of data with respect to aggregation and chunking is recordings of spoken language. Unlike previously addressed sources, these are often produced by multiple "authors" (speakers), and in some cases, they come with purely artificial "text" boundaries, corresponding to whenever the recording device happened to be turned on and off (the arbitrariness of these boundaries is similar to segmentation according to pages in a book). Two notions of homogeneity can be pursued here, and a choice must be made, depending on research priorities. One option is to go with *situational homogeneity* and extract continuous spans of the transcript as chunks, irrespective of how many speakers they might involve, preserving the dynamics of interaction. The other one is to aggregate per-speaker threads in each transcript and chunk those, if they get too long, which leads to *authorial homogeneity*, similar to the treatment of short web posts.

We ended up taking the second path, our rationale being the following: we had a fairly varied sample of multi-party speech data, from broadcast discussions of various kinds and field interviews to informal private conversations. Conventionally, these would be associated with different types of linguistic devices, even more so in the case of a language such as Czech, where borderline diglossia sharpens the contrast between formal and informal communication. Crucially though, many speakers have an imperfect command of the formal variety of Czech and some even eschew it deliberately. Inasmuch as it reflects an important aspect of variation in the language, this is a phenomenon we wanted to capture and indeed succeeded in doing so: the highest scoring chunk on the second dimension, which we interpret as high spontaneity (see Section 5.2), is a TV appearance by rabbi Karol Sidon, who is well known for not worrying too much about adhering to the formal convention; runners-up in spontaneity all come from private conversations. In the situational homogeneity approach, Sidon's idiosyncrasies might have been averaged out, and in any case, the mix of speaker threads would have been hard to disentangle and interpret.

Throughout this section, we have been appealing to an intuitive notion of what "texts" are, i.e. books, chapters, articles, posts and other structures defined by simple formal criteria associated with a particular medium, as they seem to be frequently used as observation units in MDAs. It should be pointed out that there are much more sophisticated definitions of "text", generally along the lines of Halliday and Hasan (1976: 23): "a passage of discourse which is coherent in these two regards: it is coherent with respect to the context of situation, and therefore consistent in register; and it is coherent with respect to itself, and therefore cohesive". However, applying this definition has a key prerequisite: a well-developed account of registers in

the language under scrutiny, rooted in both extratextual and intratextual criteria, because identifying a text relies on establishing consistency of register. Unfortunately, in the first iteration of an MDA, as is our case, establishing a more nuanced view of registers is precisely one of the goals of the analysis, not a given.

Furthermore, the practical value of the definition is questionable because it controls in no way for the length of the resulting units. This is problematic in the context of MDA: as we have seen previously, linguistic variation is operationalized in a way that makes it sensitive to the length of the entity on which it is being measured, frequency relativization notwithstanding.

But even if the length issue was avoided, Halliday and Hasan's more refined approach to texts presents additional challenges, stemming from the fact that texts are now entities whose boundaries are not readily available in the source data but must first be established. This in itself is tricky[5] because for both humans and machines, identifying linguistic structures gets harder and increasingly ambiguous the more semantically and pragmatically involved these structures are.

2.2 Heterogeneity of corpus

A welcome side effect of working with text excerpts is that given a target size, a corpus of excerpts is likely to be more diverse, simply by virtue of containing more observation units (i.e. chunks), than a corpus of full texts. Since some of the linguistic features to measure may require manual tagging or error correction, it is an advantage to keep the overall size of the corpus manageable, without compromising too heavily on diversity. But having more units is only half the story; the other is how to pick them judiciously.

Given this was the first, bootstrapping iteration of Czech MDA, we started from a purely extratextual set of diversity criteria, based on available metadata. This is why in the web portion of the corpus, the sub-classification is rather coarse-grained: such metadata were simply not at hand. For instance, recent

5 It should be noted that e.g. Egbert and Schnur (2018) give a less skeptical view on methods of automatic and manual delimitation of texts as observation units. However, their final position on this matter is rather pragmatic: "Ultimately, the segmentation of spoken language into operationalised texts must be achieved with careful consideration of what is appropriate and useful given the aims of a particular study".

research by Biber and Egbert (2016) has shed light on the different types of blogs in English; unfortunately, such an endeavor has yet to be replicated on Czech. In order to ease future references to different parts of the corpus, Table 1 gives an overview of the structure of Koditex: three modes of communication ultimately subdivide into 45 classes of texts, aiming at roughly 200,000 words per class, subject to data availability. A detailed account of the corpus, including acknowledgments of data sources and annotation tools, is available at https://wiki.korpus.cz/doku.php/en:cnk:koditex.

For almost all categories, we had more data than needed (the exceptions being *pri* and *adm*). The way we selected the chunks to include in the final corpus can be described as *diversified stratified random sampling*: we sampled each class separately (stratified sampling), while paying particular attention to within-stratum diversity. For this purpose, we used additional metadata available about the individual text chunks in the respective classes.[6] In order to be included as a source of diversity, a metadata field had to fulfill two criteria:

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/ dialog)	201,690	82
			inf (informal unprepared private dialog)	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

(*continued*)

6 For instance: position of chunk in the original text (beginnings, middle portions or ends), name of periodical, translation vs. original text (in *wri*); topic category (in *wik*); gender, age, education (in *inf*), etc.

Table 1: (*continued*)

Mode	Division	Superclass	Class	Tokens	Text chunks
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 (trade journals)	fts (formal and technical sciences)	210,010	71
			hum (humanities)	207,916	69
			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 & magazines)	lei (leisure)	hou (crafts & 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**	**3,292**

Note: Token counts exclude punctuation.

1. neither too few nor too many levels (at the extremes, the communication channel is the same for all chunks within a given class, IDs are all different; neither are thus useful for diversification)
2. at least somewhat plausible as a source of linguistic variation (e.g. sequential numbering of speakers in a conversation was excluded)

The chunk selection algorithm[7] then proceeded as follows: within each class, the first chunk was picked at random. It was then compared to all remaining unselected chunks using a trivial dissimilarity measure (simply a tally of the disagreements between the designated diversity metadata fields of the candidate chunk and all previously included chunks). The chunk that scored highest was included next. This operation was repeated until the prescribed quota for the given class was reached, or failing that, until exhaustion of the candidate pool.

While sharing some properties of stratified random sampling, this method goes beyond that. In stratified sampling, each stratum (or class in our case) is defined by some core criteria, and to fill a given quota, no other criteria are considered. Our method, on the other hand, takes into account all relevant metadata within each stratum, not only those defining the stratum, and strives for a maximally diverse sample while respecting the overall design of the corpus with predefined proportions for each class.

3 Features

The pivotal point of an MDA is the list of linguistic features which can reasonably be expected to vary in different registers and in terms of which language variation is therefore explored and analyzed. We devised ours based on previous literature in the field, starting with obvious Czech counterparts to the original English features in Biber's seminal book (1988), then consulting Czech grammars and stylistics handbooks,[8] but also more narrowly focused articles and monographs,[9] and ultimately relying on our own intuition as native speakers. The list spans the major levels of linguistic description: phonology, morphology, lexicon, syntax, text and pragmatics. Each feature is operationalized as a corpus query to be run against

7 We are indebted to Jiří Václavík for fleshing out this algorithm.

8 E.g. Hoffmannová et al. (2016); Čechová et al. (2008); Mistrík (1989); Petr et al. (1986); Karlík et al. (1995); Cvrček et al. (2010).

9 E.g. Kodýtek (2008); Čermák (2014); Miller and Weinert (1998); Čmejrková and Hoffmannová (2011).

Koditex or a custom automated extraction procedure in cases where we hit the limitations of the corpus query language. A small amount of manual cleanup was performed in a post-processing step where both necessary and feasible. The first draft of the list contained 160 features which were gradually trimmed down to 122 for various reasons.[10]

Alongside traditional relative frequency-based characteristics, we also decided to include as separate features the size of the inventory (number of types) of pronouns, prepositions and conjunctions. In general, a larger inventory arises by having recourse to less common representatives of a given category, which is a partial indicator of lexical richness, especially in a text of small size (up to 5,000 words), but each of these type-based features also had individual motivations, contrasting with respective frequency-based features. For instance, a high frequency of pronouns indicates strong situational rooting and/or cotextual linking; the number of *different* pronouns tells us how rich and varied these links are. The use of many different types of prepositions in a text follows the need to convey complex and manifold relations between notions. An extensive inventory of conjunctions is motivated by similar concerns with respect to relations between sentences or phrases and can be expected e.g. within academic discourse.

Raw inventory size obviously correlates with text length, so it will not do. However, simply dividing by the text length is not enough; the resulting type-token ratio (TTR) still exhibits the correlation. Therefore, we opted for zTTR, a properly standardized version of TTR (Cvrček and Chlumská 2015). zTTR is based on comparing the observed TTR with a distribution of reference TTR values gathered from a population of texts of identical size. Inspired by the *z*-score, this index allows us to compare the size of type inventories regardless of text length. As can be seen from Table 2, type-based features ended up playing a significant part in some of the factors (cf. their communality in the factor model), and their correlations with chunk length are indeed fairly small.

As less part-of-speech-specific indicators of lexical richness, we also introduced features based on the zTTRs of word form unigrams and lemma bigrams, complemented by a measure of lexical uniformity, Yule's K characteristic (Oakes 1998: 204), which is also conceived as text length insensitive.

10 Twenty-three features proved difficult or impossible to operationalize (low precision or recall or both or general intangibility). Some of these involved homonymic markers that can only be disambiguated based on syntactic or semantic context, which is tricky to formalize; eight were exclusive to written language; seven were too sparsely distributed, occurring in less than 15 % of the chunks.

Table 2: Correlation between type-based features and chunk length.

Feature	Spearman's corr. with chunk length	Feature communality
Type inventories:		
→ Pronouns	−0.103	0.736
→ Prepositions	0.147	0.451
→ Conjunctions	−0.024	0.492
Non-specific lexical richness indicators:		
→ Word unigrams	0.139	0.89
→ Lemma bigrams	0.102	0.778

Note: The last column shows the feature communality which is often interpreted as the reliability of the feature and represents the share of feature variability explained by all factors.

4 Number of dimensions

Dimensions are established through FA, the statistical procedure by which related features are grouped into factors (latent variables), which are then interpreted as dimensions of variation. The number of dimensions to extract is one of the input parameters of the FA, which leads to somewhat of a chicken and egg paradox: one of the goals of the MDA procedure is precisely to determine the number of dimensions of variation in the language under scrutiny, but prior knowledge of that number is required in order to perform the FA in the first place.

Biber (1995: 120) states that "there is no mathematically exact method for determining the number of factors to be extracted". Revelle (2017: 38) adds (omitting a good number of the methods he lists):

> Each of the procedures has its advantages and disadvantages. [...] The scree test is quite appealing but can lead to differences of interpretation as to when the scree "breaks". Extracting interpretable factors [i.e. the number of factors which yields the most plausible interpretation] means that the number of factors reflects the investigator's creativity more than the data. [...] The eigen value of 1 rule, although the default for many programs, seems to be a rough way of dividing the number of variables by 3 and is probably the worst of all criteria.

In our case, the scree plot (Figure 2) breaks possibly as early as the fourth factor, which would suggest a three-factor solution, and certainly no later than the tenth or thereabouts. The "eigenvalue of 1" rule yields no less than 19 factors

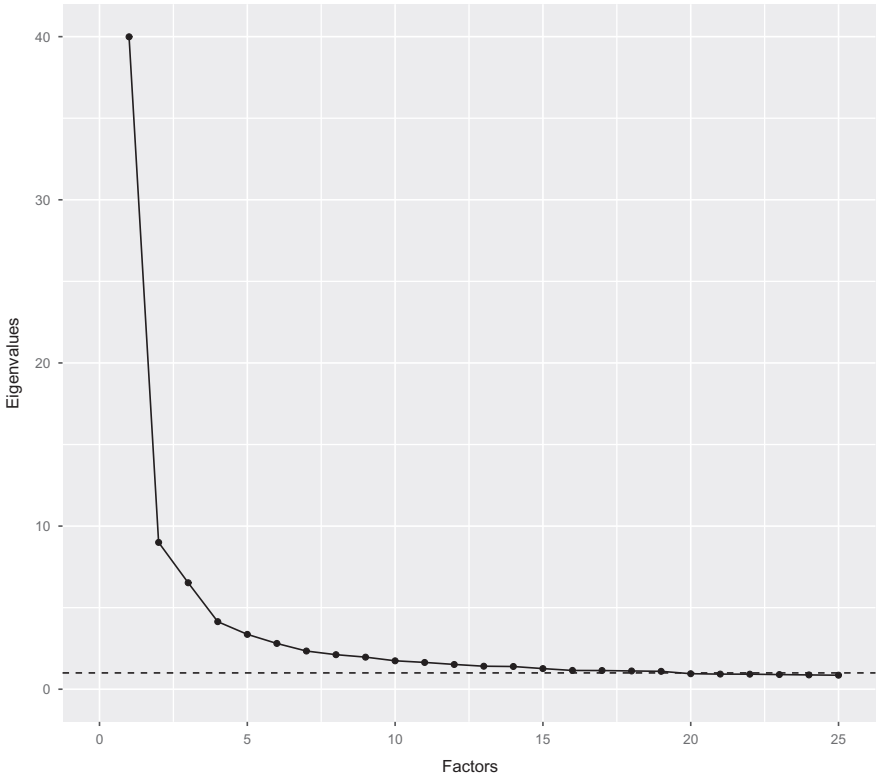

Figure 2: Scree plot of factor eigenvalues. Dashed line corresponds to eigenvalue = 1.

with eigenvalues > 1. In the light of previously performed MDAs of other languages, which had 4–9 dimensions, neither 3 nor 19 appears very plausible.

As for the "extracting interpretable factors" option, this essentially amounts to manually comparing the merits of different models and is a fairly popular choice when performing FA in the context of MDA. The original MDA of English considered several solutions ranging from four to eight factors (Biber 1995: 121) and finally settled on seven factors as the most appropriate model. Unfortunately, in our case, we found ourselves unable to reach a convincing conclusion in this way. No one model emerged as clearly dominant, i.e. as yielding the most satisfying interpretation.

However, the underlying notion of comparing several acceptable models and selecting the best (or least bad) one feels sound. We therefore set out on an attempt to come up with a quantitative formulation of the comparison criteria we were struggling to apply when carrying out the manual interpretations.

4.1 Quantifying a "good" model

Taking a hint from previous MDAs, we considered four factors to be the smallest model worthy of interpretation (the three-factor solution would not do as it mixes together too many tendencies, which makes it almost impossible to interpret). Beyond ten factors, the FA procedure did not complete successfully due to the violation of various computational constraints. We thus ended up with seven competing models, from 4 to 10 factors. Our goal is now to assess the relative merits of the models and settle on one of them.

Let us simplify the problem a little bit to make it easier to reason about: think of dimensions as groups of strongly correlated linguistic features which should ideally represent a recognizable axis of variation. This is not entirely accurate because it would mean that the association between dimensions and features is binary (as in, either a feature belongs to a dimension or not), whereas, in fact, it is gradual and polar. In other words, each feature is associated with all the dimensions, just to different degrees, as expressed by its loadings[11] on said dimensions, and the loadings can be positive or negative, corresponding to the opposing poles of the dimension. Still, this simplification is a reasonable approximation of how dimensions are usually interpreted: all features whose absolute loadings exceed a certain threshold (usually 0.3) are considered as "belonging" to the dimension in question and contributing towards its functional interpretation.

Now, suppose for a moment that we already *know* – from an independent source – what the true dimensions of variation are, i.e. how the features should ideally be grouped. Would that help us in choosing the best among our seven empirically derived candidate models? Indeed it would, we could just compare the seven models with the one true solution and pick the model which groups features in the same way that the reference grouping does.

This idea is illustrated in Figures 3 and 4. In each figure, a hypothetical correspondence between an ideal solution (left) and an empirical model (right) is visualized (with lines being individual features). The first one (3) shows the tidy case, where the correspondence between the groups is perfect; the second (4) shows a messy, tangled case, where the correspondence is more or less random.

The problem of selecting the best among our seven available models can therefore be restated as picking the one whose correspondence with the ideal solution is least tangled. In order to do so, however, we have to get rid of our

11 In FA, loadings are real numbers between –1 and 1 which quantify the correlation between a feature and a particular dimension.

Figure 3: Example of a "tidy" relationship between the "true" grouping of linguistic features into dimensions (on the left) and an MD model (= groups of features inferred via FA, on the right).

simplifying assumptions (binary association between features and dimensions, independent knowledge of true solution) and come up with a way of measuring "tidiness".

First off, concerning the true solution: we have obviously no way of knowing that. What we do have is the candidate models, as different empirically derived perspectives on what it might be. By combining these perspectives and noticing the commonalities, we can build an approximation designed to bring out the similarities between them – tendencies which are so strong that most of the models agree on them, which means they have a good chance of reflecting the true state of things.

Such an approximation can be built by clustering the features using hierarchical cluster analysis (HCA). The input data for HCA consists of a table with features in rows and their absolute loadings on all dimensions across all empirical models in columns. Based on this table, HCA examines the similarities between features. For instance, it may be the case that whenever feature A loads heavily on a dimension, B does too, whereas C systematically tends to load on

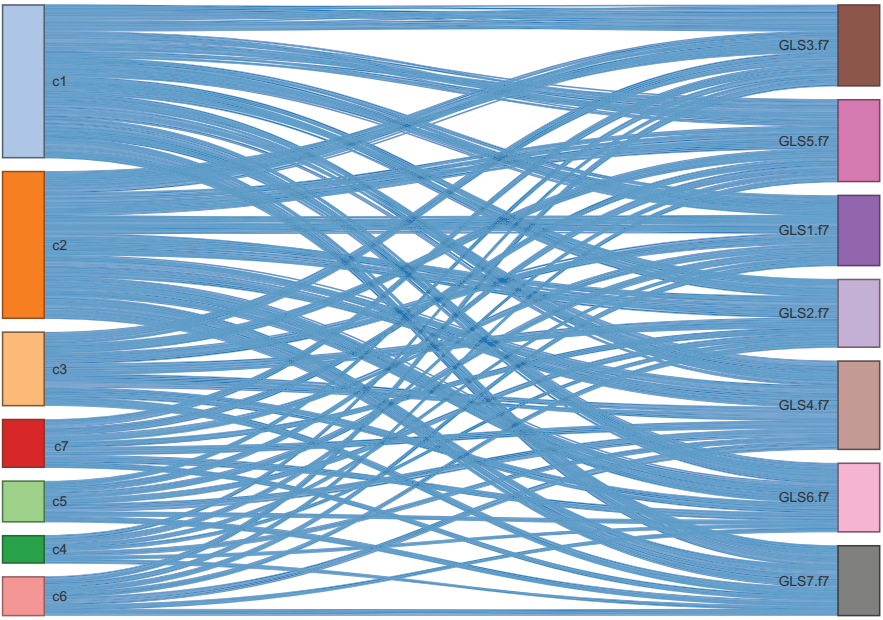

Figure 4: Example of a "tangled" relationship between the "true" grouping of linguistic features into dimensions (on the left) and an MD model (on the right).

different ones. As a result, in the output of HCA, A and B will be represented as closer to one another than either to C because this is something all or most of the models agreed on. Based on this proximity information, the HCA output can then be partitioned into different numbers of clusters. This means we can generate 4- to 10-cluster partitions, compare them with the 4- to 10-dimensional models by calculating each pair's mutual tidiness, and pick the model which looks best across all comparisons. In the absence of a clear winner, we can at least narrow down the pool; this is after all another heuristic aid, not an exact procedure.

To get an idea of what this might look like, skip ahead to Figure 6. Each cell is a tidiness score characterizing the correspondences between an n-cluster partition and an m-dimensional model, for n, m from 4 to 10. A comparison of the models with the hypothetical "ideal" solution would be a similar chart with just one row. However, since we are only approximating this solution, it seems fairer to generate a cluster partition of the same complexity (i.e. same number of groups) for each of the candidate models and take into account all possible comparisons (i.e. a model's performance across its entire column) when picking the winner.

Now, how do we measure this tidiness we keep talking about? Another way to think about tidiness is, how much does knowing one grouping (an *n*-cluster partition) tell us about another grouping (an *m*-dimensional model)? Take a look at Figures 3 and 4: in the tidy case (3), we can predict the membership of a feature in the right-hand groups based on its membership in the left-hand groups with perfect accuracy. In the tangled case (4), we might as well toss a (7-sided) coin.

Information theory provides us with a measure which quantifies precisely this, i.e. the amount of information that can be obtained about one random variable through another random variable: mutual information (MI).[12] However, as can be seen from Figure 5, we must also account for the total amount of information (or joint entropy) associated with the two groupings of features: if the groupings are complex and encode a lot of information, then there is a lot of opportunity for some of it to be mutual. In other words, such complex systems are likely to have a fairly high amount of MI in *absolute* numbers, but when compared to the total amount of information in the system, the *proportion* can be quite small. This is why we define tidiness as the proportion of information in the system that is mutual:

$$\text{Tidiness} = \frac{\text{Mutual information}}{\text{Joint entropy}}$$

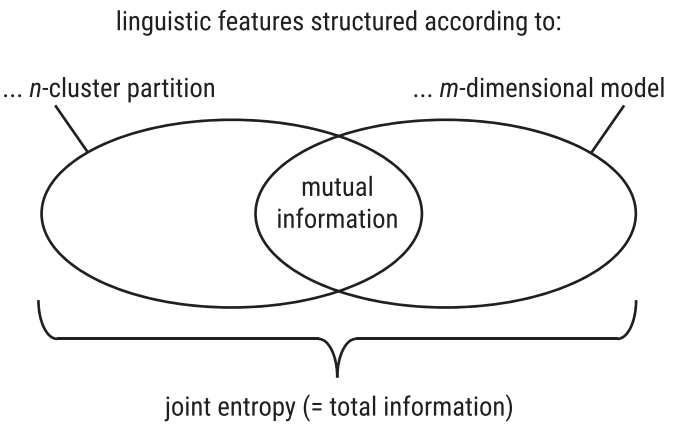

Figure 5: Comparing two groupings of features using information-theoretic measures: mutual information and joint entropy.

12 *Pointwise* mutual information is widely known as an association measure used to quantify the collocation strength (MI-score). To clear up a potential misunderstanding: our use of the concept has nothing to do with collocations.

Rephrased in the language of Figures 3 and 4, tidiness quantifies how tidy the relationship between two groupings is, while keeping in mind how tangled it could get in the worst case. As a bonus, the use of these information-theoretic measures also allows us to obviate the last remaining obstacle, which is non-binary association between features and dimensions: both MI and entropy work with probabilities, so they handle this out of the box, provided that we interpret loadings (their absolute values, to be precise) as probabilistic weights on the strength of association between features and dimensions.

The details of how to actually calculate tidiness are somewhat mathematically involved, and going over them here in an unambiguous, approachable and replicable way would disrupt the flow of the article. We have therefore decided to put them in a separate online document, together with a reference implementation in R and runnable examples; cf. https://github.com/czcorpus/mda.

The output of the method on our Czech data is given in Figure 6. The highest tidiness across comparisons with different cluster partitions emerges in the column representing the eight-dimensional model. In all rows but the last (representing the four-cluster partition), the value in this column surpasses all other values, yielding the highest column total, which suggests that the eight-dimensional model is least prone to split features which belong together on the one hand and merge dimensions which should remain separate on the other. This was therefore the model we adopted.

5 Interpretation

The input data[13] set for the FA consisted of the scores achieved in the 122 features by the 3,292 chunks. The scores were *z*-normalized within each feature. The FA itself was performed in R (R Core Team 2017) using the *fa* function from the *psych* package (Revelle 2017). Oblique rotation (*promax*) was chosen because some correlation between dimensions was expected and confirmed *ex post*. Conceptually, this translates to an assumption that registers involve bundled choices across several dimensions. The factoring method we settled on was *generalized weighted least squares* (hence the dimension numbers GLS1 to GLS8), which offered the best performance with respect to avoiding Heywood cases. The amount of variance in the data explained by individual factors and their cumulative share is summarized in Table 3. The total figure is 56 % (the sum is in the last column of the second row), which is comparable to the 52 % reported by Biber (1995: 121).

13 The entire data set is available via the TROLLing repository (doi: 10.18710/QAJKZW).

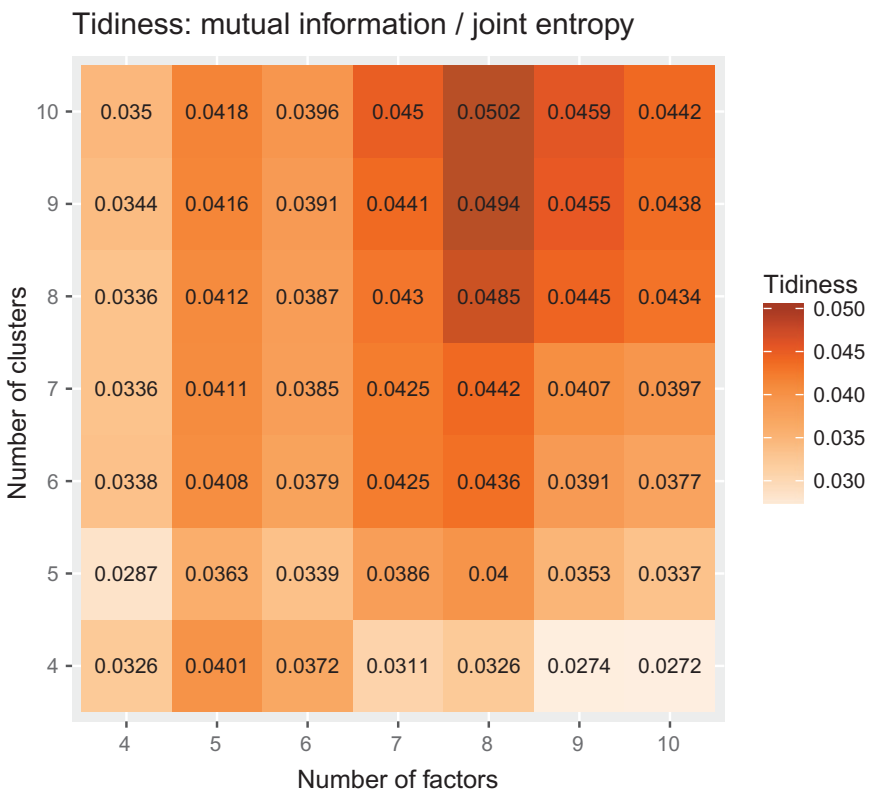

Figure 6: Tidiness, i.e. mutual information of models with a given number of clusters (rows) and factors (columns), divided by their joint entropy. Darker (higher) is better.

Table 3: Summary of the FA 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

As mentioned previously, there is no way to determine the optimal number of factors unequivocally, FA being an exploratory method. There is only guidance in the form of heuristics such as the ones described above. Ultimately, the criterion of success is whether the model yields satisfactory insights into the

mechanisms underlying linguistic variation, which is undeniably subjective to a certain extent. For completeness' sake, we summarize below the key aspects of our interpretation of the results.

5.1 Dimension 1: Dynamic (+) vs. static (–)

Similarly to MDAs of other languages, the first dimension, which explains the largest proportion of shared variance, indicates a difference in favoring verbal vs. nominal constructions (for comparison to MDAs of other languages, see Section 5.9). Among the features with high positive loadings are verbs as a part of speech and several verbal subcategories. Another group of features which positively correlate with this dimension consists of function words (pronouns and conjunctions) and adverbs. Features which load strongly on the negative pole are primarily nouns and adjectives in various functions. Other salient features include secondary prepositions, word length, passives and thematic concentration (Popescu et al. 2007: 68).

This division of features hints at opposing text composition strategies: a speaker or writer either concentrates on the elaboration of clause members (the static pole) or proceeds "faster" to their communication goal by adding new clauses (the dynamic pole). On the other hand, features with loading around 0 suggest that this dimension is indifferent to the preparedness of speakers/writers: they are either directly associated with spontaneous spoken communication or with lexical richness.

The distribution of text categories along this dimension reveals two shades of "verbality": chunks with the highest positive scores are not only narrative (e.g. various kinds of novels) but also reflective (as evidenced by verbs of thinking in private correspondence or web forums). The other side of the scale is dominated by information-dense chunks from official documents, hard science papers and encyclopedias. Overall, this dimension clearly separates academic literature and professional journals on the negative side from fiction, private correspondence and web interaction (Facebook, forums) on the positive side. This suggests that the difference between verbal and nominal discourse is also associated with the difference between subjective and objective perspective.

5.2 Dimension 2: Spontaneous (+) vs. prepared (–)

The second dimension essentially reflects differences in the conditions under which a given piece of discourse is put together. The dominant contrast here is

between the written and spoken modes. While writing allows one to refine turns of phrases *ad nauseam*, spontaneous spoken language is online production and perception (cf. Auer 2009) under heavy time constraints. Speakers are forced to juggle tradeoffs between eloquence, intelligibility, timeliness, propriety, etc.

In terms of features, the spontaneous end of the spectrum is strongly linked with contact expressions, fillers and also pronouns, which can otherwise be dropped in many contexts in Czech, so their explicit use signals a trend towards redundancy of information. This is confirmed by features such as word reduplication, Yule's K and an increased tendency to resort to formulaic language. While these may sometimes be a flaw of online production, redundancy is first and foremost a feature which makes successful communication over a noisy channel possible.

A specific type of feature occurring in spontaneous Czech is non-standard forms of the Common Czech variety (cf. Sgall et al. 1992). While these have long-established spellings used, e.g. in private communication or often in fiction, they have never been admitted into the standard and are still stigmatized in more formal settings, even in speech, leading to the aforementioned conjectures regarding diglossia. We reserve definitive judgment on the subject, but the isolation of the *Spoken interactive* box in the *Dimension: GLS2* facet of Figure 7 certainly is striking. The other two boxes with a positive (spontaneous-like) median coincide with private correspondence and Facebook/forum posts.

By contrast, prepared monologs, though delivered in speech, lie on the opposite end of the spectrum, along with administrative texts, economic news, Wikipedia articles, etc. These all exhibit a higher incidence and larger type inventory of prepositions, clauses with interrogative or relative adverbs, lexical richness (zTTR) and a large amount of nouns and longer words in general.

5.3 Dimension 3: Higher (+) vs. lower (–) level of cohesion

The third dimension marks a difference in how often connecting devices and means of intratextual reference are used. The highest positive loadings are associated with features like relative clauses and pronouns. The tendency towards a larger type inventory of conjunctions, another positive feature, suggests the need to express a broad spectrum of semantic relationships between clauses and entails frequent use of verb forms. The only feature which loads negatively on this dimension are numerals (incl. numbers).

In other words, dimension 3 informs us whether discourse elements (facts, ideas, opinions) are intricately linked together, in order to spell out explicitly

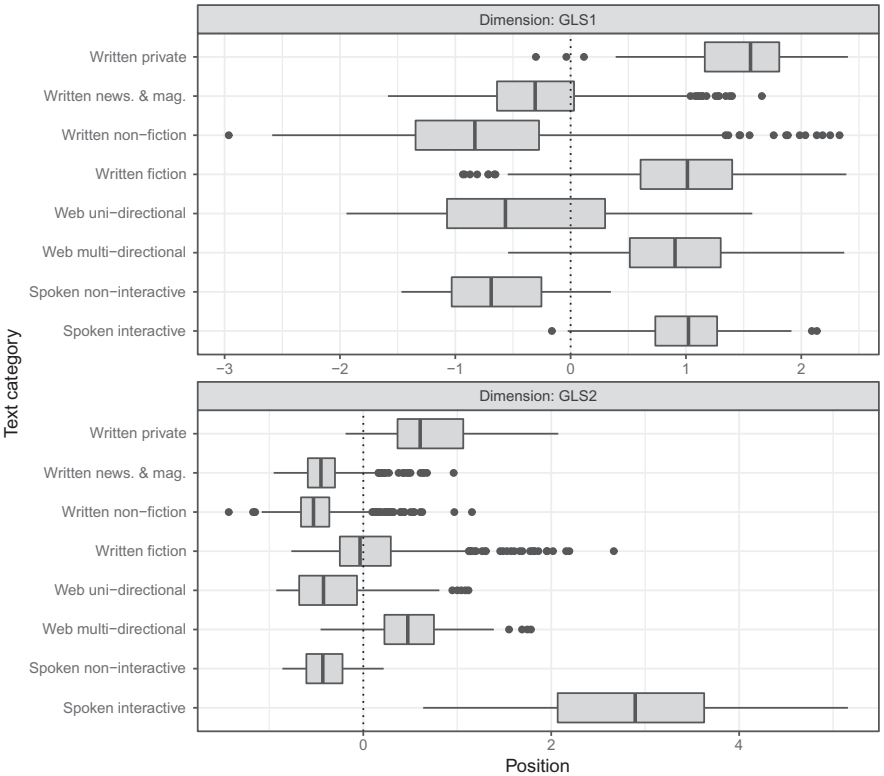

Figure 7: Position of text chunks with respect to dimensions 1 and 2. Chunks are grouped by their division attribute in the Koditex corpus (cf. corpus description in Section 2.2) to indicate global tendencies of the categories relative to the dimensions.

how they form a coherent whole, or whether this coherence is left implicit (perhaps because it is considered obvious) or achieved using linguistically unsophisticated structuring. Rudimentary structuring in the form of juxtaposition, whose only embellishment is occasional numbering, is one reason for the high incidence of numerals on the negative pole; the other reason is the point about implicit because obvious coherence: numbers as content are often seen as speaking for themselves.

This view is supported by the distribution of text categories along this dimension: written-to-be-spoken speeches, social science papers and broadcast discussions (+) vs. encyclopedias and Wikipedia, informal conversations and hobby magazines (–). Whereas neither encyclopedias nor informal conversations tend to contain involved argumentation, texts associated with a

high level of formality (like presidential speeches) or with abstract entities (like sociology or law texts) focus more on how facts are interwoven into a coherent whole. A noteworthy detail is that this is the dimension which separates soft and hard science papers most noticeably, perhaps because, as hinted above, numerically expressed relationships substitute for linguistic ones.

5.4 Dimension 4: Polythematic (+) vs. monothematic (–)

This dimension particularly emphasizes differences in lexical richness. Positive loadings include the normalized sizes of different type inventories (see Section 3), while the negative side prominently features thematic concentration and Yule's K, which expresses repetitivity. This can be interpreted as a contrast between polythematic texts because spanning multiple topics requires having recourse to a larger part of the lexicon, and focused, monothematic texts. On the positive side, additional loadings include toponyms; on the negative one, a variety of mostly grammar-oriented features, which are often used within monothematic genres: verbal nouns, passives, abstract nouns, are associated with academic writing; modifiers and prepositions allow for detailed descriptive treatment of a topic.

In terms of text categories, the polythematic end of the spectrum is occupied by journalism which focuses on short and variegated pieces: several types of newspaper sections and magazines. As noted in Section 2.1, these are categories where chunk boundaries can be expected to encompass several of the original texts, as individual pieces often tend to be shorter than 2,000 words. This is undoubtedly part of the reason why these chunks exhibit higher topic variation. On the other, negative extreme, we find administrative or academic texts and professional journals, which are usually strictly concentrated on one topic. This divide may also explain a secondary contrast between concreteness (toponyms, +) and abstractness (abstract lexemes, –).

5.5 Dimension 5: Higher (+) vs. lower (–) amount of addressee coding

The positive pole brings together questions, verbs in the second person sg. and pl., the future tense, the vocative case and the imperative mood. At first glance, these seem like features typical of dialog, as opposed to monolog. The only negative feature above threshold is average sentence length in tokens, which in

theory resonates well with the notion of monolog, but in this particular case, seems to be an unreliable indicator.[14]

The negative end of the spectrum should therefore mostly be characterized by the absence of features with positive loadings, most of which point towards the presence of and interaction with a communication partner.[15] The distribution of classes in the *spo* mode corroborates this: elicited speech, which mostly consists of extended answers to interview questions, falls on the negative end of the scale (monolog), and informal private dialogs, which leverage explicit turn-taking mechanisms embodied in the features listed above, on the positive end.

However, as can be seen from Figure 8, the most extremely positive chunks along this dimension are not even spoken, they belong to fiction: screenplays, poetry, sci-fi, crime, romance, fantasy and short stories. This may come across as less of a surprise if we accept that exaggeration of typical features is part of artistic representation: the natural characteristics exhibited by real-life dialogs will tend to be overemphasized in fictional ones. Take the example of the vocative, a case designed to address partners in communication. In a real, embodied multi-party interaction, the fact that a given utterance is addressed to a particular party is often sufficiently conveyed by body language or situation.

5.6 Dimension 6: General/intension (+) vs. particular/extension (–)

While the positive end of this dimension correlates with language units which are either semantically bleached or outright function words, the negative side is associated to a large extent with features representing concrete information. The corresponding typical text categories form a relatively heterogeneous mix at first glance: encyclopedias, poetry/lyrics, fantasy novels and hobby magazines (positive) vs. sport magazines, screenplays, economic news, tabloids and news reports (negative).

The uniting quality of text categories on the negative pole seems to be that they are anchored to a particular context, admittedly fictional in the case of screenplays, but very much specified in terms of time, place and participants.

14 As suggested by the two most prominent negative text categories, *wik* and *eli*: Wikipedia articles often include itemized lists only separated by commas, and the sentence splitting in the elicited speech transcriptions was very sparse (and at any rate, artificially imparted by linguists).
15 The only exception is the future tense, which plays a complementary role: it is often recruited alongside addressee coding for linguistic tasks such as planning.

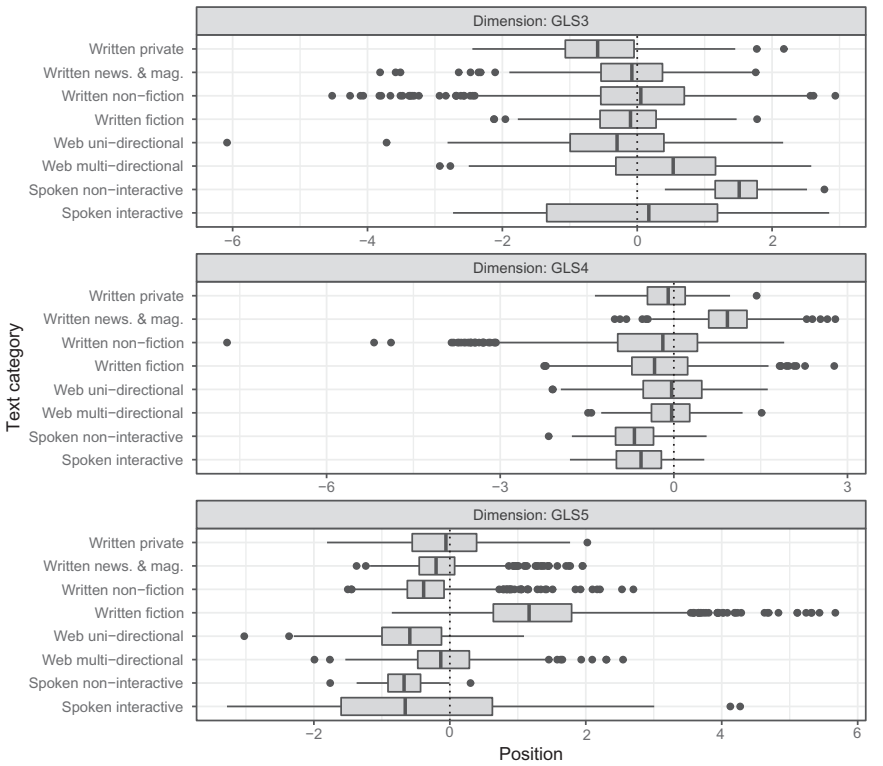

Figure 8: Position of text chunks with respect to dimensions 3–5.

This train of thought resonates with the associated features, three of which refer to named entities (names and surnames, time expressions, toponyms). This pole therefore emphasizes particular referents of words, i.e. their extension.

By contrast, the lack of named entities on the positive pole suggests that spatiotemporal anchoring is less prominent there; such situationally independent discourse can afford to be more syntactically elaborate (coordination, conjunctions) and is more likely to address less specific qualities (bleached adjectives). The emphasis is now on the intension of words, as leveraged by decontextualized encyclopedia descriptions and do-it-yourself (DIY) hobby magazine recipes. Some types of lyrical poetry can similarly be seen as eschewing the particular in favor of the universal. As for the fantasy novels, the most probable explanation unfortunately seems to be deficient feature operationalization which failed to take into account made up proper nouns in these genres.

5.7 Dimension 7: Prospective (+) vs. retrospective (–)

The primary contrast in this dimension is between the present and future tense on the positive side and the past tense on the negative. This is what guides our interpretation and the choice of labels, prospective and retrospective, with the important caveat that "prospective" has a broader meaning here. It designates discourse anchored in the present and looking out into the future, but also timeless, generic discourse, with no restricted time frame of validity, which means it is indefinitely relevant. Unlike dimension 1, which is articulated around an opposition between verbs and nouns, across part of speech categories, dimension 7 is clearly based on a distinction within the category of verbs because finite verbs on the whole, as an aggregate feature, are inert with respect to this dimension.[16]

Prospectiveness is further linked with features such as the imperative mood and the second person, which indicates that forward-facing discourse typically involves other communication partners. When discussing future, hypothetical events, a necessity often arises to specify their epistemic/deontic status, which translates into a higher frequency of lexemes expressing modality.[17] The text categories most associated with this type of language are elicited dialogs and web forums. On the other hand, retrospectiveness unsurprisingly dominates in fiction, where it manifests as narration. This is corroborated by additional features on the negative pole like the third person and possessive adjectives and pronouns, which refer to protagonists of the narrative.

5.8 Dimension 8: Attitudinal (+) vs. factual (–)

The positive extreme correlates with features related to:

- the function of particles: downtoners, restrictors, intensifiers and text organizing particles
- adverbials and conjunctions (including concessive and adversative connectives)
- modality across part-of-speech categories

The presence of these features points to texts containing careful evaluation of quality or relations together with an attempt to present the speaker's point of view. Coordination (clausal or phrasal) is the only feature with negative loading

16 Independence of the first and seventh dimensions is further witnessed by their week positive correlation ($r = 0.25$).

17 Note that unlike English, Czech does not use modal auxiliaries to express the future.

beyond threshold, but on its own, it offers no straightforward interpretation due to its versatility: it is used as a substitute for complex syntactic expressions (especially in online production) as well as for enumeration of objects in a list.

In terms of text categories, the positive extreme is occupied by web discussions and forums (usually containing one question and many answers in which participants present their viewpoint), private correspondence, elicited speech (monologs of speakers opining on a given topic) and blogs; the most "positive" class within the written mode is popular humanities non-fiction (cf. Figure 9). All these texts can be characterized by a tendency to overtly mark the message as representing the speaker's opinion or stance.

The negative extreme is populated by administrative texts, Wikipedia articles, screenplays, poetry/lyrics and tabloids. A trait common to many of these texts is the presence of enumeration (lists of names or things) and the lack of hedges or explicit stance markers.

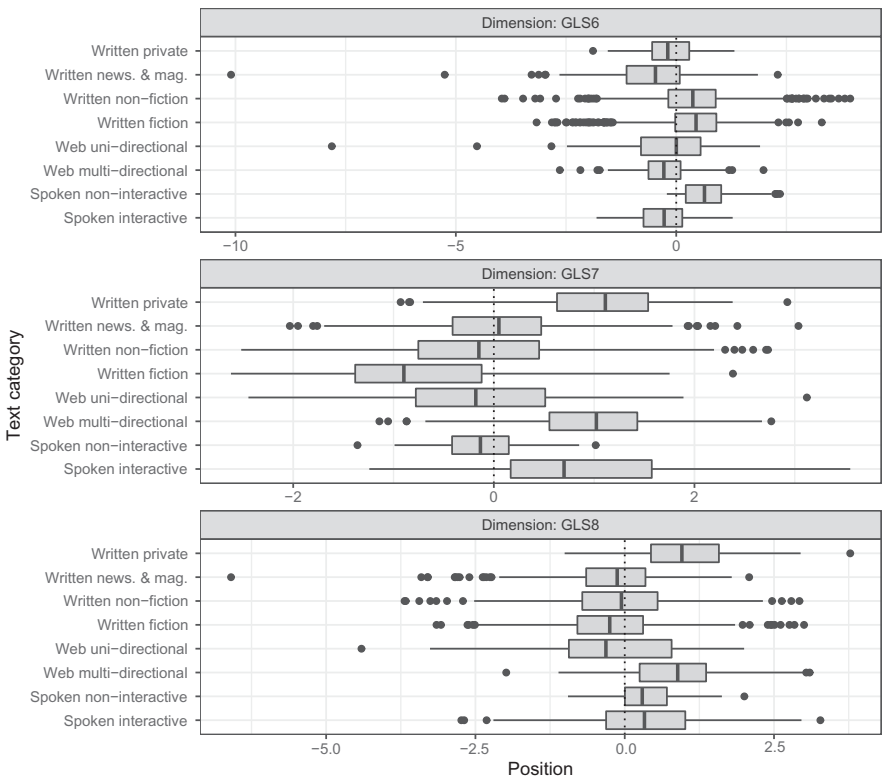

Figure 9: Position of text chunks with respect to dimensions 6–8.

5.9 Discussion

Even though a thorough comparison of MD models of different languages requires detailed analysis beyond the scope of this paper, a few observations are worth mentioning. The model of Czech supports Biber's proposal of universal dimensions (Biber 2014): both the opposition between clausal vs. phrasal and narrative vs. non-narrative discourse has emerged (the respective dimensions are GLS1: dynamic vs. static and GLS7: prospective vs. retrospective). Biber's comparison of several MD studies shows that the first dimension is usually characterized by the opposition between verbal and nominal features. Given that the most prominent text category on the "verbal" extreme of the scale is almost always private conversation, the features are functionally associated with personal involvement, interactivity and/or online production (Biber 2014: 16). Our first dimension (GLS1) is slightly different in this respect: the clear verbal vs. nominal distinction 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 (therefore the more general label "dynamic"). The aforementioned dimension focusing on narration *per se* (GLS7) is based, consistently with Biber's findings, on the contrast between past and present tense, conspicuously separating fiction from the other text categories.

As Biber (2014: 27) points out, every MDA uncovers idiosyncrasies of the target language or in the application of the MD procedure. The specific sociolinguistic situation of Czech is highlighted by dimension 2 (spontaneous vs. prepared), which 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. As mentioned above, the influence of interactivity and online production circumstances is usually visible right in the first dimension of MD models, so it is a distinct possibility that it was the presence of Common Czech features in the language situation which is reflected by our model.[18]

Another reason for the emergence of specialized dimensions is the specifics of the MD procedure. A possible example of this effect is our dimension 4: polythematic vs. monothematic, which separates newspaper articles of various

18 A conceptually similar case occurred in Spanish MDA where several features related to the irrealis mood formed an additional dimension separating prototypically spoken and written text categories alongside the verbal-nominal first dimension (cf. Biber et al. 2006).

kinds from the rest of the text categories. Undoubtedly present in other languages as well, such a distinction only manifests itself upon the inclusion of various language-agnostic measures of lexical richness (unigrams and bigrams on the positive pole, Yule's coefficient and thematic concentration on the negative) and type-based features.

6 Conclusion

Apart from the quick overview of the first Czech MDA just given, this paper's primary aim was to discuss methodological issues encountered in bringing it to life. We acknowledge that we mostly backed our decisions up with thought experiments and deductive arguments, definitive proofs being impractical.[19] Our position should therefore not be misconstrued as claiming that other solutions to these problems are wrong: we are acutely aware that our claims are tentative. Our point is rather that working within a well-established paradigm, which MDA undoubtedly is, comes with the luxury of being able to take pause at every step and explore it in depth. While the reader may reach different practical conclusions based on the theoretical considerations outlined in this article, giving these considerations thought is a luxury he or she can and should afford.

In terms of methodological innovations, we have argued that diversified stratified random sampling of text excerpts instead of entire texts is a useful approach when conducting an exploratory MDA whose primary aim is to establish the dimensions of variation in a given language. Since information about registers in Czech is still a matter of future research, it is reasonable to reflect solely extratextual characteristics in the corpus design and establish intratextually informed categories in follow-up steps on the basis of MDA results.

We have also introduced novel type-based features which leverage zTTR normalization to reduce correlation with text length to a minimum. As evidenced by the interpretation of our resulting MDA model, these features have proven instrumental in shaping some of the identified dimensions.

Last but not least, we have sketched our attempt to put the model comparison procedure (which leads to the selection of the number of dimensions to interpret) on a more formal basis using clustering and information-theoretic measures. As a heuristic decision aid, this method effectively tries to give a systematic quantification of the type of observations a linguist would hopefully consider while performing a manual qualitative comparison of models. Therein lies its promise over

19 We are thinking of ways of empirically testing some of these ideas in follow-up work.

existing generic statistical indicators. Whether that promise shall be fulfilled depends on whether the approach will prove its worth and perhaps be refined in future research.

Acknowledgements: This study was supported by the ERDF project "Language Variation in the CNC" no. CZ.02.1.01/0.0/0.0/16_013/0001758. The authors would like to thank both reviewers for valuable input.

References

Auer, Peter. 2009. On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences* 31. 1–13.

Bermel, Neil. 2014. Czech diglossia: Dismantling or dissolution? In Judit Árokay, Jadranka Gvozdanović & Darja Miyajima (eds.), *Divided languages? Diglossia, translation and the rise of modernity in Japan, China and the Slavic World*, 21–37. Dordrecht: Springer.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5(4). 257–269.

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

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14(1). 7–34.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, Douglas, Mark Davies, James K. Jones & Nicole Tracy-Ventura. 2006. Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora* 1(1). 1–37.

Biber, Douglas & Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics* 44(2). 95–137.

Čechová, Marie, Marie Krčmová & Eva Minářová (eds.). 2008. *Současná stylistika* [Contemporary stylistics]. Prague: Nakladatelství Lidové noviny.

Čermák, František. 2014. Lexis in spoken and written language. In František Čermák (ed.), *Jazyk a slovník. Vybrané lingvistické studie* [Language and dictionary. Selected studies in linguistics], 299–304. Prague: Karolinum.

Čmejrková, Světla & Jana Hoffmannová (eds.). 2011. *Mluvená čeština: hledání funkčního rozpětí* [Spoken Czech: In search of the range of its functions]. Prague: Academia.

Cvrček, Václav & Lucie Chlumská. 2015. Simplification in translated Czech: A new approach to type-token ratio. *Russian Linguistics* 39(3). 309–325.

Cvrček, Václav, Vilém Kodýtek, Marie Kopřivová, Dominika Kováříková, Petr Sgall, Michal Šulc, Jan Táborský, Jan Volín & Waclawičová Martina. 2010. *Mluvnice současné češtiny 1* [A grammar of contemporary Czech 1]. Prague: Karolinum.

Cvrček, Václav, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková & Adrian J. Zasina. Forthcoming. *Variabilita češtiny: multidimenzionální analýza* [Variability in Czech: A multi-dimensional analysis]. *Slovo a slovesnost*.

Egbert, Jesse & Erin Schnur. 2018. The role of the text in corpus and discourse analysis. In Charlotte Taylor & Anna Marchi (eds.), *Corpus approaches to discourse. A critical review*, 158–170. New York: Routledge.

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

Hoffmannová, Jana, Jiří Homoláč, Eliška Chvalovská, Lucie Jílková, Petr Kaderka, Petr Mareš & Kamila Mrázková (eds.). 2016. *Stylistika mluvené a psané češtiny* [The stylistics of spoken and written Czech]. Prague: Academia.

Karlík, Petr, Marek Nekula & Zdenka Rusínová (eds.). 1995. *Příruční mluvnice češtiny* [A reference grammar of Czech]. Prague: Nakladatelství Lidové noviny.

Kodýtek, Vilém. 2008. Variace v mluvené češtině v Čechách: sonda do ORAL2006. [Variation in spoken Czech in Bohemia: Exploring the ORAL2006 corpus]. In Marie Kopřivová & Martina Waclawičová (eds.), *Čeština v mluveném korpusu*, 132–141. Prague: Nakladatelství Lidové noviny.

Kodýtek, Vilém. Unpublished. A translation of Biber's three-dimensional model of English into Czech. https://www.korpus.cz/biblio/2722

Lee, David Y. W. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5(3). 37–72.

Miller, Jim & Regina Weinert. 1998. *Spontaneous spoken language*. Oxford: Clarendon Press Oxford.

Mistrík, Jozef. 1989. *Štylistika* [Stylistics]. Bratislava: Slovenské pedagogické nakladateľstvo.

Oakes, Michael P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Petr, Jan, Miloš Dokulil, Karel Horálek, Jiřina Hůrková & Knappová Miloslava. 1986. *Mluvnice češtiny 1* [A grammar of Czech 1]. Prague: Academia.

Popescu, Ioan-Iovitz, Karl-Heinz Best & Gabriel Altmann. 2007. On the dynamics of word classes in texts. *Glottometrics* 14. 58–71.

R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.r-project.org/.

Revelle, William 2017. *psych: procedures for personality and psychological research v1.7.8*. Evanston: Northwestern University. https://cran.r-project.org/package=psych

Sgall, Petr, Jiří Hronek, Alexandr Stich & Ján Horecký (eds.). 1992. *Variation in language. Code switching in Czech as a challenge for sociolinguistics*. Amsterdam & Philadelphia: John Benjamins.

Zasina, Adrian J., David Lukeš, Zuzana Komrsková, Petra Poukarová & Anna Řehořková. 2018. *Koditex (A corpus of diversified texts)*. Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University.

Bionotes

Václav Cvrček

Václav Cvrček (PhD Charles University Prague, 2008) is an Associate Professor of corpus linguistics at Charles University, Prague. His main research interests are corpus linguistics methodology, quantitative linguistics, and corpus-assisted discourse studies.

Zuzana Komrsková

Zuzana Komrsková holds an MA in Czech language and she is currently a PhD student at the Institute of Czech Language and Theory of Communication, Charles University, Prague. Her special areas of interest are Internet communication, corpus linguistics, and corpora of spoken language.

David Lukeš

David Lukeš holds an MA in Phonetics and English & American Studies; he is currently a PhD student at the Institute of the Czech National Corpus, Charles University, Prague. His research interests include Czech phonetics, corpora of spoken language, and quantitative methods in linguistics.

Petra Poukarová

Petra Poukarová holds an MA in Czech Language and Literature; she is currently a PhD student at the Institute of the Czech National Corpus, Charles University, Prague. Her research focuses on Czech phonetics, corpus linguistics, and corpora of spoken language.

Anna Řehořková

Anna Řehořková holds an MA in Czech language and she is currently a PhD student at the Institute of the Czech National Corpus, Charles University, Prague. Her academic areas of interest are diachronic linguistics, Czech subordinate clauses, and corpus linguistics.

Adrian Jan Zasina

Adrian Jan Zasina holds an MA in Slavic Studies and he is currently a PhD student at the Institute of the Czech National Corpus, Charles University, Prague. His main research interests include data-driven learning, gender & language, and corpus linguistics.