A translation of Biber's three-dimensional model of English into Czech

Vilém Kodýtek

Abstract: A simple procedure was employed to estimate three Biber's functional factors (here labeled *formalness*, *stance*, *narration*) in a Czech corpus consisting of 2443 texts from six registers, namely informal conversation, stance-focused formal speech, personal correspondence, fiction, periodic press and academic prose. 28 variables (linguistic features) were used. All data come from the Czech National Corpus. The result resembles that of Biber (2004) for a diversified corpus of English conversation. In both cases the functional factors are only weakly correlated with each other. It is shown that a substantial correlation between the factors is related to the functional (genre) imbalance of a corpus. An assumed difference between corpora of Czech informal conversation was confirmed and elucidated.

Key words: Czech language, spoken Czech, corpus linguistics, functional dimensions, factor correlation, corpus balance

1. Introduction

In my study of variation in spoken Czech (Kodýtek 2007ab), I compared corpora from different places in the Czech Republic. Related to it are the questions of *functional* balance and *functional* similarity of the corpora. There is also the question of what level of generality if any the results represent. I decided to carry out a multidimensional analysis which makes it possible to treat simultaneously a set of data related to my work. As there had been no such study of Czech, it appeared useful to include written registers as well.

2. Earlier work

Factor analysis ("FA") (eg. Oakes 1998, 96ff) is an intermezzo between a researcher's choice of a corpus and variables and his or her final interpretation of the factors. Referring to the theoretical foundation of the latter, Lee (2000) speaks of a *wow*! criterion. Nevertheless, FA has proved an efficient tool in various sciences incl. linguistics.

Biber (1987) applied FA to a diversified English corpus, a model of English language, and interpreted the factors in terms of communicative functions (summary in Biber 1993). He was able to explain 52 % of variance using seven factors, of which the first ("A" below) accounted for 27 %. The first four factors, ranked in order of decreasing significance, Biber interpreted as follows:

A) information-focused v. involved production,

B) narrative focus,

C) elaboration v. situation dependent reference,

D) overt expression of argumentation/persuasion.

Later, in a diversified corpus of English conversation, Biber (2004) found three dimensions corresponding approximately to the dimensions A, D, and B above (in this order). Whatever complex the dimensions are, for practical purpose it is useful to assign simple labels to them. I will use 1 *formalness* (or F), 2 *stance* (S), 3 *narration* (N). Biber found these three factors universal in terms of applicability to different types of corpora (e. g., Biber 2004, Biber et al. 2004) and even different languages (Biber 1995).

Lee (2000) observed that diversified corpora are governed by the dichotomy written v. spoken which may obscure subtler variations specific to particular registers. Of the above mentioned 1987 dimensions, the written v. spoken dichotomy is related to A and C. Moreover, C is correlated significantly with dimensions A, B and D, and thus appears redundant, at least from the technical or pragmatic point of view. However, Lee says that the narrative dimension (B) is functional in fiction only while D (argumentation/ persuasion) is instable. From his own corpus, composed of 66 written and spoken genres, he extracted via FA four factors, of which only two had an interpretation: lexical and syntactic written v. spoken, respectively. These two factors are significantly correlated.

Flowerdew (1993) observed the spoken versus written opposition of linguistic features in a corpus composed of texts and lectures of a particular scientific discipline. Xiao and McEnery (2005) applied the 1980s Biber scheme to a corpus composed of conversation, speech and academic prose. Of the factors mentioned above, only A and C, i. e., those associated with the spoken-versus-written dichotomy were significant. They showed that essentially the same picture can be obtained in much a simpler way via the key-word analysis, thus in fact confirming the mutual dependence of the first two Lee's factors. It is not clear from their paper, however, whether the picture is the same also in details, such as the distribution of texts within particular registers.

Gries (2006) proposed a robust method of analysis of (individual) linguistic variables which essentially (though not entirely) eliminates the wow! criterion. In his two examples, he has demonstrated, among other, that the hierarchic structure of texts in a corpus (a kind of its multidimensional representation) can be observed using just one variable. His results do not contradict with those of Biber. It is more the other way round: the former resemble – at least qualitatively – the latter: The

3. Analysis

Corpus. In what follows, I will discuss my corpus in terms of six registers, namely 1) informal conversation, 2) formal speech, 3) personal letters, 4) fiction,[1] 5) periodic press[2] and 6) academic prose,[3] and nine subcorpora characterised in Table 1. However, an element of the corpus in the analysis is a *text* as specified below. My choice of what does *text* mean in a particular register is a reasonable compromise between what I might have chosen had it been feasible and the levels of granularity in the Czech National Corpus ("CNC") where all my data come from.

[1] *Fiction* includes both Czech authors and translations from other languages.

[2] *Press* consists of daily newspapers, local newsletters and professional, arts & culture, economic, political, entertainment and other magazines.

[3] *Academic prose* consists of textbooks and miscellaneous-purpose publications. The latter segment is composed of various genres, such introductions into, summaries of and popularisations of various sciences, journals of professional bodies, as well as original work on subjects related to the Czech culture, such as Czech history, arts and literature studies, linguistics etc.

[Table 1]

CNC is a collection of written and spoken corpora at the Institute of the Czech National Corpus of the Charles University in Prague. The parts of CNC related to this work and the definitions of *text* in each of them are:

a) Spoken part includes

- three collections of informal (private) conversation: from the city of Prague, the city of Brno and the largest Czech region – Bohemia; *text* = all utterances of a speaker in a file; in the Prague and Brno corpora, however, different speakers of the same gender and grade of education belonging to the same age group are not distinguished in a file and, hence, are included here in one *text*.
- two collections of unprepared formal speech, i. e., monologues of speakers who (simply speaking) were made to think that they were responding in an opinion pool concerning specific issues, such as education, women's position in the society etc; *text* = file (= speaker).

b) Written unprinted part consists of personal letters from all over the Czech Republic; *text* = file (= letter).

c) Written printed (SYNchronic) part consists of two genre-diversified corpora SYN2000 and SYN2005 and a corpus limited to periodic press (SYN2006pub). In the present study included are:

- Fiction; *text* = text (with an opus.id code)
- Academic prose; *text* as above
- Press (newspapers & magazines); *text* = a volume (year) of a title.

Only texts with at least 500 words were included and each of them was given equal weight in the analysis.

Variables. I started with 43 variables (linguistic features), then eliminated insignificant items and, where possible, combined overlapping low-frequency items. The final list of 28 input variables is given in Appendix 1.

For on-line CNC search I used the Bonito software provided by the CNC Institute. The SYNs are morphologically tagged, which I fully utilised except for checking and correcting results for less regular (low-frequency) phenomena. Collecting data in the remaining, untagged corpora involved much more manual work but I could utilize some of data collected earlier.

Approach and calculation. I bypass the formal procedure of FA for several reasons. First, the time for the wow! criterion to be applied came at the first glance at the *z*-cores of variables by register where the 2004 Biber dimensions could clearly be seen. Second, this is an introductory analysis with rather a simple set of variables. I was not interested in any set of variables (possibly strongly correlated and with no or doubtful linguistic interpretation) but in My goal is not to determine any variables but more or less independent set of variables. Third, I am interested in interpretable dimensions rather than a very formal analysis which may be of little interpretation value (see part 4).

I performed the calculation in a Microsoft Excel file as follows:

In terms of the six registers' *z*-scores by variable, my choice of dimensions is as follows: *formalness* ("F") (conversation < speech < press (<) academic),[4] *stance* ("S") (speech > other registers) and *narration* ("N") (fiction > other registers). Except for a weak sign of a fourth dimension, *scheduling* (personal correspondence > other registers, with the future tense and time reference as variables), I was not able to identify any other apparent structure at the level of registers.

Assigning a weight of +1 or –1 to each significant variable and zero to all others (for each dimension), I obtained a row of 28 weights for each dimension. Then multiplying the weight row with each row in the z-score matrix, I obtained my "factors" for each of the 2443 texts.

4 Results

The result obtained for the corpus is depicted in Figure 1, where points are centroids of subcorpora and circles are "median spheres", within which 50 % of texts are located. Additional information is given and comparison with Biber (2004) is made in Table 2. Factor loads are summarized in Appendix 2.

[4] In case of prepositions, the order is .. academic < press.

[Figure 1]

[Table 2]

5 Discussion and conclusion

In terms of dimensions F, S and N, taking into account both the centroid position and the median sphere, the Bohemian (ORAL2006) and Brno (BMK-N) conversation corpora are similar, while the Prague corpus is somewhat different.[5] There is some difference in the positions of the centroids of the two corpora of formal speech, too. However, their median spheres highly overlap, which is why I consider them similar and, hence, convenient for the purpose of comparison of regional features. Essentially no difference was found between Bohemian and Moravian subcorpora in the corpus of personal correspondence KSK-dopisy.

Let us have a short look at how the factorial solution discriminates within registers.[6] In most texts in the Bohemian corpus of conversation the speakers oscillate between focus on stance, context and telling. However, if we look at outliers in the register of conversation, the contact-oriented v. information-oriented opposition, stance or argumentative v. context-focused opposition as well as narrative focus are apparent. Extreme S scores in academic prose appear in line with overtly argumentative v. descriptive opposition.[7]

Lee (2000) has shown that (a) Biber's model of the English language was arbitrary to some extent and (b) his results were not strictly reproducible. It is very important to know that there are two highly correlated spoken vs. written dimensions,[8] e. g., for developing devices for automated analysis of texts. However, it tells us little about what methodology to use for the analysis of the particular registers. On the other hand, Biber's model

In general, two functional dimensions need not be perpendicular to each other and often they are not. However, there is no reason (apart from the resercher's choice of texts or

[5] The difference is big enough to be observed. In my study of the Brno and Prague spoken corpora (Kodytek 2007), I noted that ORAL2006 (which was published when I was finalising my study) "appears to me more intimate than the PMK-N and BMK-N." It turns out that I was able to observe a difference between ORAL2006 on one side and the two city corpora (as a whole) on the other, however, not that between the PMK-N and BMK-N, which I studied simultaneously. The source(s) of the difference has yet to be searched for.

[6] I distinguish between *genre* (text type) – a functionally homogeneous collection of texts in a corpus, with funcional criteria external to the language, and *register* – a variety of language associated with a genre (text type). The reason is that it may find (and it expectionally did in my analysis) that some outliers are incorretly classified as belonging to the genre.

[7] A textbook of process law and a textbook of the international rules for work in chemical and biological labs score highest while textbooks of economic geography and of biology of cell have the lowest scores.

[8] At least in some models of Czech one would most probably find a morphological written vs. spoken dimension, too. However, in this analysis, both formal and informal variants

genres) for any pair of factors F, S and N to be significantly correlated, neither.[9] Both stance and narration can be formal as well as informal. A stance need not have anything to do with narration or can be expressed as a narrative. Indeed, there is no significant correlation between the three factors in question, neither in Biber (2004), nor in the present study, and both could probably be made orthogonal without a significant change in the distribution of texts along the axes. I hypothesize that each significant correlation of the factors should have an explanation in terms of the composition of the corpus or register.

[Table 3]

The correlation between the factors in the registers is displayed in Table 3. Conversation, speech, letters and fiction conform with the hypothesis, the two genre-diversified ones, press and academic prose, do not. Unfortunately, most genres in these two registers are not large enough to be measured. Nevertheless, I was able to identify a source of correlation in press and academic prose in Table 3. It is related to the genre balance in the register (cf. *textbooks* in Table 3). Low correlation between Biber's factors thus turns out to be a marker of the functional balance (or homogeneity) of a corpus or register.

Fig. 2 Change of factors in time in newspapers by titles

To summarize:

1) Using a "translation" of the Biber model into Czech, I was able to confirm and elucidate a fine difference between corpora of informal conversation in Czech.
2) My result for the Czech corpus compares well with that of Biber (2004) from his diversified conversation English corpus, both in terms of comparable linguistic features and in terms of the linear dependence of the factors.
3) I have shown that low correlation between factors F, S and N indicates the functional balance of the corpus.

[9] *Some* correlation has to be expected.

REFERENCES

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

Biber, D. (1993): Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics* 19, 219-241.

Biber, D. (1995): *Dimensions of register variation. A cross-linguistic comparison*. Cambridge University Press: Cambridge – New York – Melbourne – Madrid – Cape Town – Singapore – São Paulo.

Biber, D., 2004, Conversation text types: A multi-dimensional analysis. *7es Journées internationales d'Analyse statistique des Données Textuelles*. Available on <www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_000.pdf.>

Biber, D, S. M. Conrad, R. Reppen, H. P. Byrd, M. Helt, V. Clark, V. Cortes, E. Csomay, A. Urzua (2004): *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. TOEFL Monograph Series 25, January 2004. Available on <www.ets.org/Media/Research/pdf/RM-04-03.pdf >.

Flowerdew, J. (1993): Variation across Speech and Writing in Biology: A quantitative Study. *Perspectives* (*City University of Hong Kong*) 5, 75-87. Source: The Hong Kong Journals Online, available at http://sunzi1.lib.hku.hk/hkjo/.

Gries, S. T., 2007, Exploring variability within and between corpora: some methodological considerations. *Corpora* 1, 109-151.

Kodýtek, V., 2007a, Mluvená čeština v Praze a v Brně: sonda do mluvených korpusů (Spoken Czech in Prague and Brno: a probe into spoken corpora). *Slovo a slovesnost* 68, 23-37.

Kodýtek, V., 2007b, Variace v mluvené češtině v Čechách: sonda do ORAL2006 (Variation in spoken Czech in Bohemia: a probe into ORAL2006). Presented at the conference Čeština v mluveném korpusu (Czech in Spoken Corpus), Prague, September 12-14.

Lee, D. Y. W, 1999, *Modelling variation in spoken and written language*: *The multi-dimensional approach revisited*. PhD Thesis. Dept. of Linguistics and Modern English language, Lancaster University.[10]

Oakes, M. P., 1998, *Statistics for Corpus Linguistics*. Edinburgh University Press, Edingburgh.

[10] See also a powerpoint presentation on <http://www.comp.lancs.ac.uk/ucrel/local/crg/dlee_phd/>.

Xiao, Z. a A. McEnery, 2005, Two approaches to Genre Analysis: Three Genres in Modern American English. *Journal of English Linguistics* 33, 62-82.

Czech National Corpus – corpora ORAL2006, PMK, BMK, KSK-dopisy, SYN2000, SYN2005 and SYN2006pub. The Institute of the Czech National Corpus of the Charles University, Prague. Available on <http://ucnk.ff.cuni.cz>.

Table 1 Corpus composition (all parts come from the Czech National Corpus)

Register	Private conversation			Formal speech	
(Sub)corpus in CNC	ORAL2006	BMK-N*	PMK-N*	BMK-F*	PMK-F*
Region/city	Bohemia	Brno	Prague	Brno	Prague
No. of words in thous.	943	270	248	204	407
No. of *texts*	471	154	140	124	196
Production date	2002-2006	1994-99	1988-96	1994-99	1988-96

Register	Private letters	Fiction	Academic prose	Press
(Sub)corpus in CNC	KSK-dopisy	[1]	[2]	[3]
No. of words in thous.	360	30 821	9 421	393 501
No. of texts	459	492	247	160
Production date	1990-2004	<2004	1990-2004	1989-2004

* PMK / BMK = the Prague / Brno Spoken Corpus; the suffix -N / -F stands for its informal / formal part.

[1] Subcorpora COL and NOV of SYN2005 (most work written >1989, some earlier). [2] SCI and TXB of SYN2005. [3] SYN2006pub + PUB of SYN2000 and SYN2005.

Table 2 Correlation of factors with selected variables in registers

	Informal	Formal	Letters	Fiction	Textbooks	Biber
	convers.	speech				(2004)
Formalness						
attribut. adject.	0.48	0.65	0.56	0.76	0.23	0.35
prepositions	0.33	0.62	0.56	0.55	0.34	0.47
relative clauses	0.53	0,58	0.52	0.59	0.18	0.45
Stance						
I think (that)	0.43	0,46	0.17	0.35	0.15	0.66
likelihood adv.	0.41	0,37	0.40	0.46	-0.07	0.43
Narration						
past tense	0.60	0,55	0.64	0.95	0.72	0.79
present tense	-0.62	-0,44	-0.64	-0.79	0.62	-0.51

Table III Correlation between factors by register (genre)

Register	F v. S	F v. N	S v. N
Press (by volume)	-0,24	-0,75	0,11
Press by title	-0,29	-0,73	-0,02
Academic prose	-0,05	-0,47	-0,05
- textbooks	-0,01	-0,11	0,18
Fiction	-0,25	-0,24	0,02
Formal speech	0,20	-0,21	-0,11
Personal letters	-0,09	-0,12	0,00
Informal conversation	0,19	0,20	-0,04
Corpus	-0.19	-0.07	-0.19
Biber (2004)	-0.24	0.09	0.08

Appendix 1: List of variables with weights

Var. No.	Variable	Weight
formalness		
1	attributive adjectives	+1
2	verbal nouns	+1
3	nouns (f) ending -*ce*, -*ost* (mostly abstract)	+1
4	prepositions	+1
5	relative clauses controlled by noun	+1
6	*to* ‘that‘/‘it‘ (to je ‘*it is*‘)	–1
7	*ale* ‘but‘	–1
8	2sg. reference [see note 1]	–1
9	*jak* ‘how‘	–1
10	1sg. present tense	–1
11	3rd person pers. pron. (*v*)*on*(*a*/*o*/*i*/*y*) ‘he, she, (it), they‘	–1
12	demonstr. pron. *ten* (lemma excl. *to* – No. 6)	–1
13	*ještě* ‘still‘(*ještě ne* ‘not yet‘) + *už* ‘already‘	–1
14	questions	–1
15	time reference [2]	–1
stance		
16	*myslím* (*si*), *že* ‘I think/mean/guess that‘	+1
17	selected nouns [3]	+1
18	attributive pronouns and indefinite numerals	+1
19	*by* + *aby* conditional „particles“	+1
20	*nebo* ‘or‘	+1
21	modal verbs [4]	+1
22	*protože* ‘because‘	+1
23	infinitive forms of verbs	+1
24	likelihood adverbs [5]	+1
narration		
25	past tense	+1
26	*se* – the clitic of reflexive verb forms	+1
27	*když* ‘when‘	+1
28	3rd pers. present tense (sg. + pl.)	–1

Notes:

[1] 2sg. imperative forms *počkej* ‘wait‘, *řekni* ‘tell‘, *hele* ‘look‘, forms of 2sg pers. pron. *tě, tebe, tobě,* contractions related to 2sg *tos, tys, ses, sis*, *abys, bys*, other 2sg-related *viď* ‘isn’t it? ‘, *seš* ‘you’re (sg., col.)‘;
[2] forms *brzo/y* ‘early‘*, čtvrtek* ‘Thursday“*, dávno, den / dn*(*ů/y*) ‘day(s)‘, *denně* ‘daily‘,*dnes*(*ka*) / *neska* ‘today‘ *době* ‘(period of) time‘,*vždy*(*cky*) / *dycky / pořád / furt* ‘always‘, *hodin*(*u/y*) ‘hour(s)‘, *jednou* ‘once‘*, konci / konec* ‘end‘*, let*(*ech*) */ rok*(*u/ů/y*) ‘year(s)‘, *letos* ‘this year‘, *měsíc*(*e*) ‘month(s), moon‘*, nedávno* ‘recently‘, *neděli* ‘Sunday‘, *někdy* / (*v*)*občas* ‘sometimes‘,*pátek* ‘Friday‘, *pololetí* ‘half a year, semester‘*, pondělí* ‘Monday‘,*pozdě* ‘late‘, *předti/ím* ‘before (adv.)‘, *příští* ‘next‘, *ráno* ‘morning‘*, sobotu*

‘Saturday‘, *sto/aletí* ‘century/ies‘ *středu* ‘Wednesday‘, *teď*(*ka*) ‘now‘, *tejden/týden* ‘week‘*, úterý* ‘Tuesday‘, *včera* ‘yesterday‘ *večer / večír* ‘evening‘ *většinou* ‘mostly‘, *začátku* ‘beginning‘*, zejtra / zítra* ‘tomorrow‘;
[3] lemmas *člověk* ‘man (human being)‘*, názor* ‘opinion‘*„ pád* ‘case‘*, pár* ‘pair‘*, pocit* ‘feeling‘*, pořádek* ‘order‘*, problém* ‘problem‘*, případ* ‘case‘*, přístup* ‘approach‘*, systém* ‘system‘*, věk* ‘age‘*, vztah* ‘relation‘*, způsob* ‘way, manner‘*, věc* ‘thing‘*, cena* ‘price‘*, doba* ‘period of time‘*,náhoda* ‘chance‘*, otázka* ‘question‘*, podstata* ‘substance, essence‘*, potřeba* ‘need‘*, pravda* ‘truth‘*, spousta* ‘a lot of (*noun*)‘*, strana* ‘page, party, hand (in the phrase ‘on one / the other hand‘)‘, *troška* ‘a little (*noun*)‘*, většina* ‘majority‘*, hledisko* ‘view‘*, slovo* ‘word‘;
[4] lemmas *moci* ‘can, be allowed to‘, *muset* ‘must, have to‘, and cond. of *mít* ‘should, ought to‘;
[5] *asi* ‘approximately, about; probably‘, *možná* ‘may be (*adv.*)‘*, patrně* ‘presumably‘*, pravděpodobně* ‘probably‘*,zřejmě* ‘obviously‘*, jistě,* (*v*)*opravdu, skutečně* ‘really, indeed, sure ‘, *snad* ‘perhaps‘*, určitě* ‘certainly, definitely‘.

Appendix 2: Loads of variables on factors [*]

		F	S	N
1	attributive adjectives	0,91	-0,32	-0,21
2	verbal nouns	0,82	-0,14	-0,32
3	abstract nouns	0,75	0,00	-0,47
4	prepositions	0,76	-0,33	-0,02
5	relative clauses	0,73	0,00	0,01
6	*it*	-0,83	0,27	-0,23
7	*but*	-0,68	0,26	-0,08
8	2sg reference	-0,67	-0,07	0,02
9	*how*	-0,62	0,11	-0,06
10	1sg present tense	-0,60	0,35	0,06
11	3rd person pronouns	-0,61	-0,08	-0,15
12	demonstrative pronouns	-0,59	0,44	-0,09
1				
3	*still/yet* + *already*	-0,64	0,02	0,16
14	questions	-0,57	-0,11	-0,03
15	time reference	-0,53	-0,01	0,12
1				
6	*I think* (*that*)	-0,11	0,71	-0,19
17	selected nouns	0,39	0,55	-0,28
18	would, in order to	-0,07	0,64	0,04
19	attrib. proun. & indef. num.	-0,09	0,56	-0,12
2				
0	*or*	-0,19	0,53	-0,34
21	modals	-0,15	0,61	-0,09
2				
2	*because*	-0,34	0,54	-0,05
23	infinitives	-0,03	0,59	0,10
24	likelihood adverbs	-0,42	0,56	-0,04
25	past tense	-0,09	-0,32	0,87
26	3pers. present tense	-0,02	0,42	-0,72
27	reflexive verbs	0,15	0,10	0,67
2				
8	*when*	-0,27	0,12	0,56

[*] See Appendix 1 for an accurate description of the entries.