Corpus as a Means for Study of Lexical Usage Changes

Michal Křen
Jaroslava Hlaváčová
Charles University in Prague

The paper presents a corpus-based method for obtaining ranked wordlists that can characterise lexical usage changes. The method is evaluated on two 100-million representatively balanced corpora of contemporary written Czech that cover two consecutive time periods. Despite similar overall design of the corpora, lexical frequencies have to be first normalised in order to achieve comparability. Furthermore, dispersion information is used to reduce the number of domain-specific items, as their frequencies highly depend on inclusion of particular texts into the corpus. Statistical significance measures are finally used for evaluation of frequency differences between individual items in both corpora.

It is demonstrated that the method ranks the resulting wordlists appropriately and several limitations of the approach are also discussed. Influence of corpora composition cannot be completely obliterated and comparability of the corpora is shown to play a key role. Therefore, although highly-ranked items are often found to be related to changes of language usage, their relevance should be cautiously interpreted. In addition to several general language words, the real examples of lexical variation are found to be limited mostly to temporary topics of public discourse or items reflecting recent technological development, thus sketching an overall picture of lifestyle changes.

1. General outline

The paper presents a corpus-based method aiming to discover usage development of individual words and its possible interpretations. In order to minimise influence of other factors, it is necessary to base the procedure on large comparable corpora. The method is evaluated on two corpora that were found to comply with this requirement and to be the most suitable for this task within the framework of the Czech National Corpus project: SYN2000 and SYN2005. Both are representatively balanced 100-million monolingual synchronic written corpora covering two consecutive time periods: SYN2000 contains texts from the 1990s, while SYN2005 concentrates on texts from the first half of the 2000s. They are disjunctive, i.e. none of the texts was included into both of them. Despite their intentionally very similar concept, the corpora differ in two important aspects that have to be taken into consideration. First, as a natural consequence of the improvement accomplished within the five years period between their publication dates, SYN2005 was processed by notably enhanced tools including tokenisation (dividing the text into sequence of tokens), segmentation (sentence boundary recognition), lemmatisation and morphological tagging, cf. Hajič (2004) and Spoustová et al. (2007: 67-74).

In addition to the lemmatisation and tagging, both corpora were manually bibliographically annotated, including also detailed information about text register and genre. The second major difference between the corpora comes up at this point: both corpora were compiled according to modified sampling criteria based on this annotation. SYN2000 contains 15% of fiction, 25% of professional literature and 60% of newspapers and magazines, while SYN2005 contains 40% of fiction, 27% of professional literature and 33% of newspapers and magazines. As reported by Králík and Šulc (2005: 357-366), the proportions are in both cases based on sociological research that evaluated text reception (reading) rather than production (writing). They also show that the sampling criteria are elaborate and fine-grained; the main text registers are subdivided into many categories thus constituting a complex set of conditions that determine inclusion of individual texts into a particular corpus. We are aware that any external manually-assigned evaluative mark-up of this kind is more or less

subjective. However, it remained consistent throughout the period of compilation of both corpora, thus allowing easy quantification of comparability of the two corpora in terms of the external mark-up with desired granularity.

There are several papers that concentrate on corpora comparison based mostly on wordlists or n-gram frequencies, e.g. Kilgarriff (2001: 97-133) or Rayson and Garside (2000: 1-6). However, they compare corpora synchronically, i.e. any detected difference is considered to indicate difference in corpora composition, while the primary aim of our comparison is to discover diachronic language development. This means that the corpora are expected to be different and that this difference should be ideally caused only by the language development itself. Although this requirement can hardly be met in practice, it makes results of the statistical methods questionable for this purpose. As Asmussen (2006: 33-48) points out, trying to keep selected linguistic features constant may obscure detection of possible diachronic changes in case the features are not invariant themselves. It is therefore reasonable to rely on the annotation scheme described above despite its possible drawbacks.

In order to minimise influence of the modified sampling criteria, frequencies of individual items in both corpora were normalised. The term normalisation in this paper refers to recalculation of the original frequency so that it corresponds to a virtual corpus where all the main text registers are equally represented, i.e. with one third share. Although the normalisation allows for comparability of lexical frequencies, Křen (2007: 109-120) shows that their simple comparison highlights domain-specific items, frequency of which depends on inclusion of individual texts. It is therefore desirable to employ some kind of dispersion measure in order to distinguish domain-specific items from general-language ones. There are various ways of how to utilize the dispersion information in addition to the frequency. The method described here is based on average reduced frequency (ARF) introduced by Savický and Hlaváčová (2002: 215-231). ARF was chosen mainly because it was well attested by Čermák and Křen (2004) as a primary classification criterion during compilation of the newest Frequency Dictionary of Czech (henceforth referred to as FDC). The ARF formula is given in Figure 1:

$$\mathrm{ARF} = 1/v \sum_{i=1}^{f} \min\{d_i, v\}$$

Figure 1: ARF formula.

where f denotes frequency of given item (type) in the corpus, d_i distances between individual occurrences of this item (tokens) in the corpus and $v = N/f$ (N is the corpus size), i.e. it is an average distance between the individual occurrences. ARF for a given item can be viewed as a correction of its frequency based on distribution of its occurrences in the corpus: the more even the distribution is, the closer ARF approaches to the frequency. On the other hand, ARF of words that occur only in a single small cluster is close to 1 regardless of their frequency. The maximum value of ARF is thus equal to the frequency (for items with all $d_i = v$), its minimum value is 1 (for items with $f = 1$). However, ARF of common function words is typically around a half of their frequency, but it is considerably smaller (typically 10 times or more) for domain-specific items that occur only in a few documents. Because ARF minimises influence of less common words with incidentally high frequency and the normalisation allows direct comparability, normalised ARF was used instead of the frequency.

Given the normalised ARF (henceforth called NARF) values for a particular item in both corpora, significance of the difference between the two values is evaluated statistically using χ^2, LL and CBF measures. This paper does not concentrate on issues regarding statistical evaluation of natural language data in general as e.g. Oakes (1998) or more specifically Rayson and Garside (2000: 1-6). To sum up the basic implications, word frequencies tend to differ across any two texts just because of non-random nature of the language. Selection of statistical measures used for the evaluation is based on results of several papers, mainly on a survey of statistical approaches that can be used for detection of words characteristic for given texts presented by Kilgarriff (1996: 33-40). It is beyond the scope of this paper to explain the

selection criteria and to describe formulae of widely-used standard measures (χ^2 and LL). Figure 2 thus shows only the CBF formula:

$$\mathbf{CBF} = \frac{\chi^2}{\sqrt{\mathbf{a} + \mathbf{b}}}$$

Figure 2: CBF formula.

where *a* and *b* denote frequencies of the given item in the compared corpora. Basic properties of the CBF can be found in Křen (2007: 109-120) who also discusses a theoretical problem related to the different nature of differences between low vs. high frequency words. For instance, let us consider Table 1 that shows three lemmata together with their NARF in both corpora:

lemma	SYN2000	SYN2005
esemeska (SMS message)	0	71
internetový (internet - adj.)	310	2075
kraj (country or county)	3093	6011

Table 1: Selected NARF differences.

It is not clear which of the NARF differences should be considered the most significant and what is thus the desired result of their statistical evaluation. It is arguably task-dependent and often individual, as statistical significance may be only loosely connected to e.g. lexicographical relevance. As a consequence, there cannot be one universally accepted measure and this fact should limit our expectations concerning the results the measures are able to provide us with. It also emphasises the importance of human intuition and common sense that should guide their interpretation.

2. Processing description

In order to achieve basic comparability and to avoid the processing differences, both corpora were processed with the same set of tools mentioned in Chapter 1, i.e. tokenisation, segmentation, morphological analysis and disambiguation. After that, they were split into subcorpora according to the three main text registers (fiction, professional literature, newspapers and magazines). ARF values for every lemma were computed for each subcorpus separately, thus constituting partial ARF wordlists. At the same time, the ARF value was normalised and the NARF was listed next to every item in the partial wordlists. Overall wordlists for both main corpora resulted from joining the partial ones, overall NARF resulted from adding up the individual partial NARFs. Items containing a digit or not containing any alphabetic character were excluded from further processing, as well as proper names (i.e. items with lemma containing an uppercase letter).

The overall NARF values for individual items in SYN2000 and SYN2005 were evaluated by the three statistical measures: χ^2, LL and CBF. Every item was ranked according to the value given by each of the measures. Since the ranks are comparable and thus easier to evaluate than exact values, the tables below list only the ranks rather than exact values next to the individual items. Finally, the automatic ranking was manually inspected in order to evaluate the results.

There is naturally an implicit limitation of any wordlist-based approach that should perhaps be mentioned here. Such an approach can be able to detect neologisms, words disappearing from the lexicon or more generally lexical units undergoing usage changes that are salient rather than subtle or too gradual. It is also virtually incapable of detecting the most common manifestations of lexical usage changes that include semantic shift, polysemy or collocability preferences, not to mention syntax or other non-lexical language phenomena. These are discussed more generally by Asmussen (2006: 33-48) who concentrates on possible diachronic interpretations of contrastive observations on various language levels and also outlines a methodology for comparative corpus-based studies. He also stresses significant impact of corpus composition on the results. However, his observations are based on smaller corpora of the same size, but rather different concept, thus containing notably different proportions of the main text registers. On the

contrary, our paper presents only a wordlist-based approach, but at the same time introduces normalisation in order to achieve comparability of the base corpora.

3. Evaluation

This chapter summarises the most important findings discovered during the evaluation stage. The evaluation is not meant to be exhaustive; its aim is to give an overall picture of possible results, implications regarding the source data, as well as advantages and limitations of the presented method. Generally speaking, the method should be viewed as giving hints rather than ready-made lists of individual lexical items. Unfortunately, the scope of this paper does not allow to extend the evaluation further to lower-ranked items and the evaluation is thus limited to the top 50 lemmata ranked according to the individual measures. Since LL and χ^2 give very similar orderings (cf. their ranks in Table 2), it was decided to leave out the χ^2-ranked table.

lemma	SYN2000 NARF	SYN2005 NARF	LL rank	χ2 rank	CBF rank
se (reflexive -*self*)	1333375	1448643	1	1	487
na (on)	829085	892343	2	2	64924
euro (the currency)	196	1903	3	3	1
internetový (internet - adj.)	310	2075	4	5	2
ale (but)	206284	230725	5	4	5372
kvůli (because of)	8078	13189	6	6	25
cz (part of internet address)	249	1662	7	9	3
když (when)	107478	123923	8	7	1480
on (he)	190846	212356	9	8	64898
náš (our)	49795	40494	10	11	337
strana (side or political party)	46949	37933	11	12	276
být (to be)	1888160	1948762	12	10	117621
kraj (country or county)	3093	6011	13	14	15
rok (year)	131161	147379	14	13	64920
mít (to have)	303826	327699	15	15	100320
www (part of internet address)	303	1522	16	17	4
pan (Mr.)	17994	12954	17	16	101
foto (photo)	1597	397	18	20	6
už (already)	99668	112348	19	18	96088
m (abbreviation)	5065	2676	20	19	27
do (into)	300685	321683	21	21	102372
-*li* (enclitic meaning *if*)	24719	19200	22	22	269
oni (they)	120338	133229	23	24	98220
v (in)	1171477	1210877	24	23	119721
s (with)	428480	452162	25	25	108471
moci (to be able to)	163109	177825	26	26	100619
dostat (to receive)	29295	35494	27	27	1578
b (abbreviation)	5719	3408	28	28	46
webový (web - adj.)	55	641	29	40	5
dítě (child)	19132	24088	30	29	656
ona (she)	101518	112489	31	30	98595
krajský (county - adj.)	921	2231	32	34	16
jako (*like* - prep.)	177897	192287	33	31	102281
koruna (*crown* - Czech currency)	12630	16669	34	32	285
že (that)	400806	422143	35	33	112930
říkat (to say)	30053	36033	36	35	5329
internet (internet - subst.)	1189	2584	37	37	26
začít (to begin)	34277	40526	38	36	64548

z (from)	407998	428624	39	38	113966
tento (this)	179441	166336	40	39	105395
který (which)	368666	387882	41	41	113913
fax (fax)	1174	343	42	43	10
hejtman (county marshal)	100	675	43	48	7
mobil (mobile (phone))	244	983	44	45	8
mobilní (mobile - adj.)	672	1714	45	44	18
médium (media - subst.)	1684	3180	46	42	34
akciový (*stock* (business) - adj.)	3754	2139	47	46	55
jestli (if)	10744	14000	48	47	665
mluvčí (spokesman)	3641	5629	49	49	113
auto (car)	6083	8579	50	50	210

Table 2: The most significant NARF differences according to LL.

lemma	SYN2000 NARF	SYN2005 NARF	LL rank	$\chi 2$ rank	CBF rank
euro (the currency)	196	1903	3	3	1
internetový (internet - adj.)	310	2075	4	5	2
cz (part of internet address)	249	1662	7	9	3
www (part of internet address)	303	1522	16	17	4
webový (web - adj.)	55	641	29	40	5
foto (photo)	1597	397	18	20	6
hejtman (county marshal)	100	675	43	48	7
mobil (mobile (phone))	244	983	44	45	8
čtyřkoalice (4-parties alliance)	34	316	93	114	9
fax (fax)	1174	343	42	43	10
kupónový (coupon - adj.)	405	63	85	101	11
com (part of internet address)	68	416	86	100	12
bosenský (bosnian)	532	105	74	80	13
eurozóna (euro area)	5	138	199	269	14
kraj (country or county)	3093	6011	13	14	15
krajský (county - adj.)	921	2231	32	34	16
hypermarket (hypermarket)	46	284	144	163	17
mobilní (mobile - adj.)	672	1714	45	44	18
kouč (*coach* (instructor))	373	1094	62	62	19
celebrita (celebrity)	68	348	132	143	20
kli (journalistic signature)	1	90	286	397	21
mediální (media- adj.)	400	1136	63	63	22
unionista (unionist)	44	260	167	199	23
esemeska (SMS message)	0	71	344	524	24
kvůli (because of)	8078	13189	6	6	25
internet (internet - subst.)	1189	2584	37	37	26
m (abbreviation)	5065	2676	20	19	27
unijní (union - adj.)	8	112	315	375	28
í (spaced text remainder)	235	41	207	230	29
sž (journalistic signature)	0	66	379	572	30
bin (part of Arabic names)	64	277	217	238	31
sč (journalistic signature)	0	50	552	773	32
logistický (logistic - adj.)	73	293	222	240	33
médium (media - subst.)	1684	3180	46	42	34
zastupitel (representative (polit.))	357	898	108	110	35
privatizace (privatisation - sub.)	2188	1082	56	58	36

tch (journalistic signature)	2	54	638	816	37
dýdžej (deejay)	1	48	661	876	38
hn (journalistic signature)	7	77	563	649	39
moderátorka (she-moderator)	59	240	281	304	40
kandidátský (candidate - adj.)	78	287	262	278	41
privatizační (privatisation - adj.)	555	196	158	164	42
čtyřkoaliční (4-parties alliance)	2	51	682	871	43
ě (spaced text remainder)	91	11	525	602	44
globalizace (globalisation)	110	361	224	236	45
b (abbreviation)	5719	3408	28	28	46
modelka (she-model)	168	485	189	197	47
marka (German mark)	1247	572	97	99	48
outsourcing (outsourcing)	11	85	596	664	49
aquapark (waterpark)	4	56	720	877	50

Table 3: The most significant NARF differences according to CBF.

Some lemmata, especially in Table 3, can be viewed as errors (e.g. *í*) or results of processing inconsistencies highlighted by the comparison. These include mainly results of different levels of corpus cleanup (journalistic signatures *kli*, *sž* should have been preferably removed, lemma *foto* often occurs as a part of contact information in text header or footer etc.) or tokenisation (*www*, *com*, *cz*—internet addresses should not have been split into parts, although their frequency increase is thus not doubted).

Apart from these, lemmata in both tables can be roughly divided into two main groups. The first group consists of period-specific words, temporary topics of public discourse or reflection of social and technical development: *euro*, *internetový*, *hejtman*, *mobil*, *bosenský*, *kraj* etc. We argue that these words are of primary interest, although they may not represent the core of language development according to the native speaker intuitions. However, the intuitions are based mostly on spoken language, while the corpora consist only of written language which is more conservative. In other words, expectations should be based on what can be inferred from the data. It also means that highlighting the processing inconsistencies and errors should not be regarded as a drawback of the measures, because they cannot be distinguished from other data on the basis of their frequency and dispersion. Taking this into account, all the CBF-ranked results in Table 3 can be considered relevant, thus sketching an overall picture of political and lifestyle changes.

The second group consists of very frequent general language items that are not present in the CBF-ranked table. These are often function words for which the NARF difference between the two corpora cannot be readily explained. It is important to mention that Table 2 includes 8 items (*se*, *na*, *být*, *v*, *s*, *že*, *z* and *který*) from the top 10 most frequent Czech lemmata according to the FDC. Overestimation of small differences between frequent items because of their statistical significance is a well-known feature of LL and in particular χ^2. However, there are also other important causes of this finding that are interconnected and more or less present in individual cases. Deeper insight into their nature can be gained by inspecting Table 4 that shows partial NARF for the main text registers in both corpora. Since all the NARF values are directly comparable, it is easy to observe the major differences by comparing the corresponding columns.

	SYN2000			SYN2005		
lemma	fiction	news	prof. lit	fiction	news	prof. lit
se (reflexive -*self*)	641962	354992	336419	666155	439763	342723
na (on)	283818	289170	256095	296586	331795	263961
euro (the currency)	5	152	37	1	1577	323
internetový (internet - adj.)	1	205	102	9	1707	358
ale (but)	107384	55239	43659	104325	84108	42290
kvůli (because of)	3094	4159	823	3721	8370	1096
cz (part of internet address)	2	115	130	2	1353	305

když (when)	66299	22878	18299	71736	32152	20034
on (he)	122717	37101	31027	128572	53356	30427
náš (our)	16238	18894	14661	14023	15018	11452
strana (side or political party)	8356	19906	18686	8216	13568	16147
být (to be)	784186	526073	577899	785626	593507	569628
kraj (country or county)	1440	1025	626	1660	3407	942
rok (year)	21468	69921	39771	22407	77591	47379
mít (to have)	124182	100989	78653	127951	120365	79381
www (part of internet address)	1	161	139	2	1196	322
pan (Mr.)	13426	3105	1462	9309	3006	637
foto (photo)	18	1333	245	30	311	54
už (already)	59758	28562	11347	58461	43129	10757
m (abbreviation)	469	1372	3222	382	690	1603
do (into)	115172	99570	85941	120500	113316	87865
-li (enclitic meaning *if*)	6717	5071	12929	4690	3568	10940
oni (they)	54749	33411	32176	56431	42400	34396
v (in)	287032	462133	422310	286963	479401	444510
s (with)	140743	141533	146203	146722	154270	151169
moci (to be able to)	56959	48619	57530	60921	55912	60991
dostat (to receive)	12058	11975	5260	13696	16370	5427
b (abbreviation)	257	2188	3273	196	919	2291
webový (web - adj.)	0	37	17	4	423	213
dítě (child)	7566	7199	4366	8956	9803	5328
ona (she)	65515	18188	17814	68096	26103	18288
krajský (county - adj.)	83	736	101	57	1961	211
jako (*like* - prep.)	76582	41821	59491	77771	49252	65262
koruna (*crown* - Czech currency)	1130	10573	926	920	14790	957
že (that)	180558	131387	88859	183420	143801	94920
říkat (to say)	17433	8816	3802	19074	13314	3643
internet (internet - subst.)	3	584	601	26	1873	684
začít (to begin)	15989	11502	6785	17779	15289	7457
z (from)	129493	146656	131847	133220	161753	133649
tento (this)	25548	57391	96500	21134	43063	102138
který (which)	98637	139957	130070	98989	144810	144081
fax (fax)	11	512	649	33	207	102
hejtman (county marshal)	35	39	24	27	598	48
mobil (mobile (phone))	31	133	78	57	859	66
mobilní (mobile - adj.)	22	397	252	51	1259	402
médium (media - subst.)	62	978	642	87	2353	738
akciový (*stock* (business) - adj.)	70	2202	1480	79	954	1105
jestli (if)	7874	2172	698	9830	3623	544
mluvčí (spokesman)	140	3277	222	111	5288	229
auto (car)	2493	2793	795	3034	4821	722

Table 4: Partial NARF in the main text registers for the lemmata from Table 2.

Even simple comparison of the NARF values for various text registers within a single corpus can bring interesting findings. For instance, two most frequent Czech prepositions *v* and *na* have unexpectedly distinct text register distribution: while the distribution of *na* is more or less even, *v* is remarkably a "non-fiction word". However, we will now concentrate on the NARF differences in the corresponding text registers between the corpora. In most cases, the most remarkable NARF difference can be found in the newspapers, occasionally supported also by the other text registers. This observation often pertains to "fiction words", i.e. items with NARF notably higher for fiction than for the other text registers. This can be explained by arguing that

written texts in general tend to be less formal nowadays and this tendency is reflected more promptly in the newspapers because fiction and professional literature are more conservative. In terms of the main written text registers, non-formal words typical for fiction infiltrate the newspapers gradually. This causes NARF increase of words typical for fiction. It is statistically significant if the words are either very frequent themselves (*se*, *ale*, *být* etc.) and/or their typicality for fiction is remarkable (*se*, *ale*, *když*, *on*, *už* etc.). This explanation is also supported by the fact that the expected NARF increase of period-specific items and neologisms (*euro*, *internetový*, *kraj*, *mobil* etc.) is caused primarily by the newspapers.

Of course, the simple fact that the NARF differences are almost always observed in the newspapers can also raise a doubt about consistency of the external mark-up. Texts regarded as fiction in earlier stages of corpora compilation might have been gradually subsumed under the newspaper text register. This is a serious objection, therefore relative frequencies of individual lemmata for the whole newspaper register (PUB) were compared to *Mladá fronta Dnes* (MFD), one of the most popular Czech newspapers. Some of the results for a period of 1992 - 2004 are plotted in Figure 3. Lemmata *ale* (*but*), *už* (*already*) and *strana* (*side* or *political party*) are frequent general language words ranked 18, 36 and 75 in the FDC and at the same time they were selected from the top 20 of Table 4. Moreover, *ale* and *už* are non-polysemous function words typical for fiction (cf. Table 4), but at the same time their NARF increase is observed only in the newspapers. In order to base Figure 3 on the largest possible data, it was extracted from all available synchronic written corpora including SYN2000, SYN2005 and also 300-million newspaper corpus SYN2006PUB. Their overall size is thus 500 millions of tokens, the newspapers constituting 400 millions.

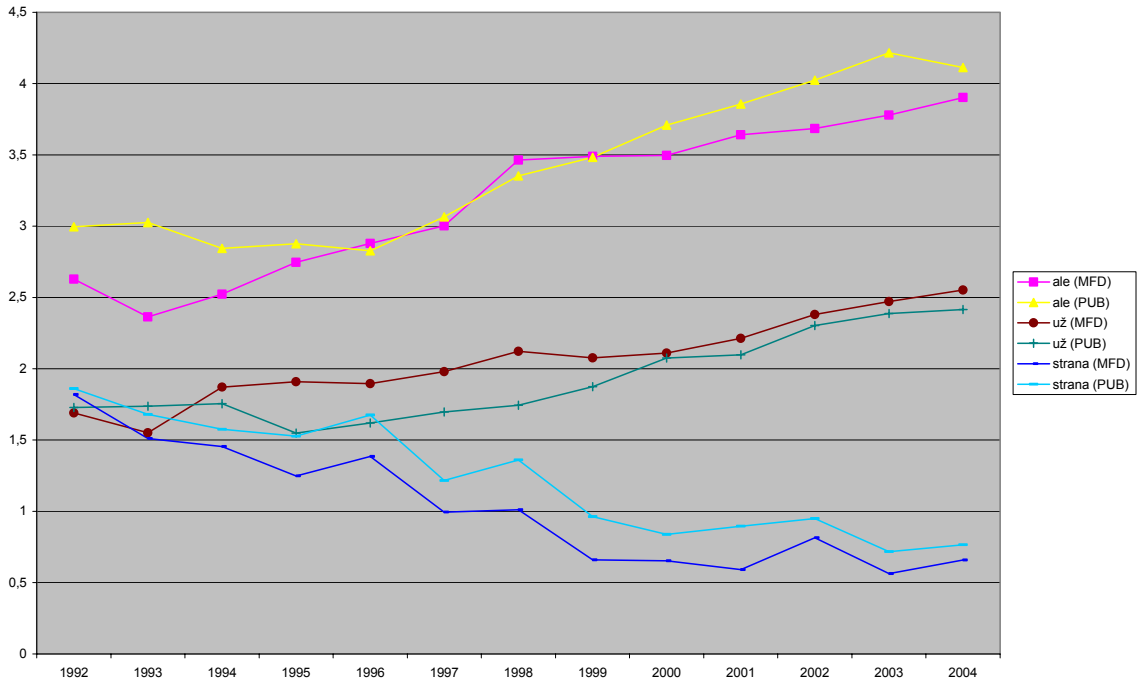

Figure 3: Relative frequency per 1000 tokens for selected lemmata

Since the usage tendencies observed in the newspapers in general are confirmed within the single newspaper title as well, the inconsistent mark-up objection should be refuted. However, there are still some unresolved issues regarding homogeneity of the data, because the repository of newspapers in SYN2000 contains titles different from those in SYN2005. Some of them did not exist at the time, some were just not available which means that single text register can be homogeneous only to some extent. Because the results for MFD and newspapers as a whole are very close, homogeneity of a single newspaper title could also be disputed. Indeed, individual issues of MFD show considerable growth in size and also gradual thematic extension within a span of ten years (ca 25,000 tokens per issue in 1992 compared to

ca 150,000 tokens in 2002) caused most probably by growing weekend supplements that thin down the original political orientation, cf. also NARF decrease of lemma *strana*. This non-homogeneity is also the most likely cause of the observed NARF increase for "fiction words" *ale* and *už* that does not correspond to native speaker intuition and can hardly be explained otherwise.

In other words, homogenous corpora suitable for diachronic comparison are scarce, because even single newspaper title may change over time considerably. Growing weekend supplements are typical for Czech newspapers in general, infiltration of popular leisure themes into originally political medium presumably influences the language in general and this is what should be measured. It may be argued that comparison of non-homogenous corpora is questionable and that well-established text-internal quantitative measures of corpus homogeneity, e.g. Kilgarriff (2001: 97-133), can be used as a source of corrective information about the nature of the data. This is a good point, although the quantitative methods are not able to distinguish between gradual changes in newspaper title composition and language development in general. In other words, full homogeneity of the base corpora cannot be primary requirement for diachronic comparison.

Apart from questions discussed so far, Table 4 shows also linguistically supported tendencies regarding the very frequent general language items. Their high NARF value seems to be a meaningful relevance criterion if supported also by NARF differences observed in all the main text registers. The most significant examples include preposition *kvůli* (*because of*), conjunction *když* (*when*) or enclitic *-li* (*if*). The latter two examples are often interchangeable, enclitic *-li* becomes rather archaic and is thus being replaced with other expressions. Table 5 shows enclitic *-li*, its potential substitutes and their variability across the main text registers that indicate some development tendencies of Czech conditional clauses.

	SYN2000			SYN2005		
lemma	fiction	news	prof. lit	fiction	news	prof. lit
když (when)	66299	22878	18299	71736	32152	20034
-li (enclitic meaning *if*)	6717	5071	12929	4690	3568	10940
jestli (if)	7874	2172	698	9830	3623	544
pokud (if)	4720	11230	10155	5568	12156	12481
zda (if)	2448	5883	3879	1962	5725	4325
jestliže (if)	1161	1608	3074	1130	1134	3480

Table 5: Partial NARF in the main text registers for selected lemmata.

However, there are also examples similar in terms of the their NARF differences that do not conform to the native speaker intuition, e.g. lemmata *dítě* (*child*), *dostat* (*to get*) and *začít* (*to begin*). Even detailed examination of their individual word forms, typical collocations and possible dispersion anomaly did not throw light on possible causes of the NARF increase shown in Table 4. Although corpus composition difference can influence even frequent items like these, it cannot be the only cause. More likely, an unexpected language development tendency seems to play a role here at least to some extent.

To mention another example, possessive pronoun *náš* (*our*) displays NARF decrease in all the main text registers (cf. Table 4). Closer investigation of its collocations revealed notable NARF decrease for right collocates *zpravodaj* (*reporter*) or *spolupracovník* (*contributor*) that mostly constitute the phrase *od našeho zpravodaje*/*spolupracovníka* (*reported by our contributor*). It is surely a reflection of corpus composition differences, but since the collocates are not very frequent, they cannot account for such a significant decrease of the possessive pronoun in itself. On the contrary, possessive pronouns in general tend to be overworked under the influence of English. Among them, only *jeho* (*his*) and *její* (*her*) display notable NARF increase in all the main text registers. The typical examples include substitution of dative of personal pronoun for possessive pronoun (e.g. *alkohol v jeho krvi zjištěn nebyl*, literally *alcohol was not detected in his blood*) or substitution of verb *mít* (to have) for verb *být* (to be) together with the possessive pronoun (e.g. *jeho šaty byly stále ošuntělejší*, literally *his clothes were more and more shabby*).

The examples are from SYN2005 and they can be more naturally reformulated as *alkohol mu v krvi zjištěn nebyl* (literally *alcohol was not detected him in blood*) and *šaty měl stále ošuntělejší* (literally *he had clothes more and more shabby*).

Because this paper is not aimed at describing Czech, the examples in the preceding paragraphs were intentionally not discussed in depth and therefore the argumentation is sometimes not complete. However, the purpose was to demonstrate that detailed survey of ranked lexical lists can direct our attention also to non-lexical language phenomena and that some of the widely-discussed language development tendencies can be confirmed by means of the observed NARF differences.

4. Conclusion

As opposed to LL and χ^2, CBF prefers lower-frequency items with greater frequency differences between the corpora to higher-frequency items with smaller differences. Although the latter are statistically more significant, observed frequency differences often require elaborate professional analysis in order to determine their cause, while relevancy of CBF-ranked results is less questionable. CBF can be thus suggested as a good choice for fully automatic detection, while the other measures give candidate lists suitable for further manual processing and interpretation that may come up with interesting findings.

The presented method was demonstrated to give appropriate results by combining normalisation with ARF instead of regular frequency. There is also a possibility to modify the desired character of the results by means of selecting a suitable statistical measure for final ranking of the wordlists. The corpus composition and homogeneity issues were focused on in order to emphasise the key role of comparability between the corpora. However, full comparability can hardly be achieved, which means that differences of various kinds can get highlighted and relevance of the results should thus be interpreted cautiously. The method is not restricted to diachronic comparison, it can be used for a variety of purposes, for instance for comparison of different variants of the same language.

Acknowledgement

Many thanks to Tomáš Bartoň for his help with computation of the individual measures.

This research has been supported by Ministry of Education grant number MSM0021620823 and Information Society grants number 1ET101120503 and 1ET101120413 by Grant Agency of the Czech Academy of Sciences.

References

Asmussen, J. (2006). "Towards a methodology for corpus-based studies of linguistic change: Contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 General Corpora of Danish". In Archer, D.; Rayson, P.; Wilson, A. (eds.) *Corpus Linguistics Around the World*. Amsterdam: Rodopi. 33-48.

Čermák, F.; Křen, M. (eds.) (2004). *Frekvenční slovník češtiny*. Praha: NLN.

Hajič, J. (2004). Disambiguation of Rich Inflection (Computational Morphology of Czech). Prague: Karolinum.

Kilgarriff, A. (1996). "Which words are particularly characteristic of a text? A survey of statistical approaches". In *Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition*. Brighton: Sussex University. 33-40.

Kilgarriff, A. (2001). "Comparing Corpora". *International Journal of Corpus Linguistics* 6 (1). 97-133.

Králík, J.; Šulc, M. (2005). "The Representativeness of Czech Corpora". *International Journal of Corpus Linguistics* 10 (3). 357-366.

Křen, M. (2007). "Variation of Czech Lexicon as Reflected by Corpora Comparison". In Levická, J.; Garabík, R. (eds.) *Computer Treatment of Slavic and East European Languages*. Bratislava: Tribun. 109-120.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Rayson, P.; Garside, R. (2000). "Comparing Corpora using Frequency Profiling". In *Proceedings of the Workshop on Comparing Corpora, Annual Meeting of the ACL Archive* 9. 1-6.

Savický, P.; Hlaváčová, J. (2002). "Measures of Word Commonness". *Journal of Quantitative Linguistics* 9 (3). 215-231.

Spoustová, D. et al. (2007). "The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech". In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*. Prague. 67-74.

SYN2000 [on-line]. Praha: Ústav Českého národního korpusu. *http://ucnk.ff.cuni.cz*.

SYN2005 [on-line]. Praha: Ústav Českého národního korpusu. *http://ucnk.ff.cuni.cz*.

SYN2006PUB [on-line]. Praha: Ústav Českého národního korpusu. *http://ucnk.ff.cuni.cz*.