M. Waclawičová, M. Křen

ORAL2008: NEW BALANCED CORPUS OF SPOKEN CZECH[1]

1. Introduction

Attention paid to spoken language has increased in the last decades, as well as its importance for linguistic research and natural language processing in general. However, compilation of spoken corpora as an indispensable source of data is very laborious and thus expensive. Nevertheless, more and more spoken corpora are being created currently. There are various approaches to their design, depth of annotation, variety of situations they cover etc.

This paper overviews corpora of spoken Czech[2] currently included into the Czech National Corpus (CNC) and their basic features, especially their balancing based on the main sociolinguistic categories of participating speakers. Finally, we will concentrate on ORAL2008, a new corpus of spoken Czech currently being prepared, its design fundamentals and automatic balancing procedure.

2. Overview of CNC spoken corpora

Currently there are following corpora of spoken Czech available for research purposes as a part of the CNC project – Prague Spoken Corpus, Brno Spoken Corpus and ORAL2006.

Prague Spoken Corpus (PSC)[3] was recorded in Prague in 1988 – 1996 and contains 819 267 tokens (including punctuation), of which 674 992 are words proper. It is built of transcriptions of recordings of

[1] This research has been supported by MSM0021620823 grant.

[2] *Čermák F., Sgall P.* Výzkum mluvené češtiny: jeho situace a problémy // SaS. 1997. № 58. P. 15–25; *Šonková J.* Mluvená čeština a korpusová lingvistika // SaS. 2000. № 61. P. 190–202.

[3] *Český národní korpus – PMK.* Ústav Českého národního korpusu FF UK. Praha, 2001. Available on-line from **http://ucnk.ff.cuni.cz**.

373 speakers in both informal and formal situations (controlled dialog with topic given in advance). Corpus annotation gives information about (in)formality of situation and basic binary sociolinguistic categories of the speakers: gender, age group (younger / older) and education (lower / higher).

There is a little prevalence of formal situations (60%), women (54%), younger speakers (52%) and speakers with lower education (53%) (cf. Table 1). However, the corpus can be considered balanced sufficiently.

Table 1. Number of tokens in selected categories in PSC

Situation	**Formal**	**Informal**	**Not specified**
	487 490	320 403	11 374
Gender	**Women**	**Men**	**Not specified**
	442 697	365 196	11 374
Age	**Younger (20–35)**	**Older (36 and more)**	**Not specified**
	423 622	384 271	11 374
Education	**Elementary/ intermediate**	**University**	**Not specified**
	436 655	371 238	11 374

Brno Spoken Corpus (BSC)[1] was recorded in Brno in 1994 – 1999 and contains 596 009 tokens (489 410 words proper). Its design stems from the PSC, it consists of transcriptions of recordings of 294 speakers in both formal and informal situations and it is balanced according to the same binary sociolinguistic categories of speakers.

Prevalence of informal situations (57%), women (58%), younger speakers (65%) and speakers with higher education (54%) is a little more remarkable here (cf. Table 2).

[1] *Český národní korpus – BMK*. Ústav Českého národního korpusu FF UK. Praha, 2001. Available on-line from **http://ucnk.ff.cuni.cz**; *Hladká Z.* Tvorba a využití korpusů češtiny na FF MU v Brně // Čeština – univerzália a specifika. 2002. № 4. P. 307–310.

Table 2. Number of tokens in selected categories in BSC

Situation	**Formal**	**Informal**	**Not specified**
	250 401	341 082	4526
Gender	**Women**	**Men**	**Not specified**
	345 142	246 341	4526
Age	**Younger (20–35)**	**Older (36 and more)**	**Not specified**
	387 499	203 874	4636
Education	**Elementary/ intermediate**	**University**	**Not specified**
	266 934	324 439	4636

The newest available corpus, ORAL2006[1], was recorded in the whole of Bohemia in 2002 – 2006 and contains 1 312 282 tokens (1 000 798 words proper). Unlike PSC and BSC, it contains transcriptions of solely informal recordings of 754 speakers. In addition to the basic binary categories, the annotation includes also exact age, education (distinguishing three grades: elementary / intermediate / university) and region of childhood residence of the speakers (according to traditional dialectological division[2]).

Table 3. Number of tokens in selected categories in ORAL2006

Situation	**Formal**	**Informal**
	0	1 312 282
Gender	**Women**	**Men**
	910 536	401 746
Age	**Younger (18–35)**	**Older (36 and more)**
	755 474	556 808
Education	**Elementary/intermediate**	**University**
	531 193	781 089

[1] *Český národní korpus – ORAL2006*. Ústav Českého národního korpusu FF UK. Praha, 2006. Available on-line from **http://ucnk.ff.cuni.cz**; *Waclawičová M.* Spoken Corpus ORAL2006, Information It Provides and General Characteristics of Spoken Text // Computer Treatment of Slavic and East European Languages. Bratislava, 2007. P. 283–289.

[2] *Bělič J.* Nástin české dialektologie. Praha, 1972; *Český jazykový atlas 1–5*. Praha, 1993–2005.

Although it was attempted to compile a balanced corpus, there was not enough material in some of the sociolinguistic categories. The table shows prevalence of women (69%), younger speakers (58%) and speakers with higher education (60%).

3. ORAL2008 and its main features

ORAL2008, a new corpus of spoken Czech sized 1 million words proper, is currently under preparation. It should be made publicly available in autumn 2008. ORAL2008 will be built from material recorded in the whole of Bohemia in 2002 – 2007 using the same repository of recordings and their transcriptions as ORAL2006. This means that the two corpora will be compatible in all respects, including also the transcription rules. However, individual recordings and transcriptions already included into ORAL2006 will not be reused in ORAL2008, so that there will be no intersection between the two corpora. Moreover, there will be two important enhancements compared to ORAL2006. First, ORAL2008 will be fully balanced according to the main sociolinguistic factors (cf. below). Second, all the transcriptions will be aligned with corresponding recordings at word-level[1].

The sociolinguistic factors[2] that strongly influence character of spoken language can be divided into two groups in relation to the needs of corpus compiling. Factors of the first group are required in only one particular realization in order to ensure authenticity of spoken language and thus constitute suitable selection criteria. Factors of the second group can occur in any realization and can be thus used

[1] *Peterek N., Kaderka P., Svobodová Z., Havlová E., Havlík M., Klímová J., Kubáčková P.* Digitisation and Automatic Alignment of the DIALOG Corpus: A Prosodically Annotated Corpus of Czech Television Debates // Proceedings of the 10th International Conference on Text, Speech and Dialogue. Springer-Verlag, 2007. P. 607–612.

[2] *Čermák F.* Mluvené korpusy // Korpusová lingvistika: Stav a modelové přístupy. 2006. P. 53–67.

for balancing of such a corpus. In practice, however, corpus is usually balanced only according to the most important of them, while the other ones can vary arbitrarily.

The first group consists of the following factors that are characteristic for authentic spoken language as opposed to written language. The key factor the other ones are related to is informality of situation. Formal situations require language with almost written character, while informal situations are bound with authentic spoken language. Private environment and unpreparedness of speech hang together with informal situation, as well as topic not given in advance and physical attendance of speakers. Relationship between the speakers is also affected, as they know each other well in such a situation. These factors suppose dialogical, not monological character of speech. In practice, they are most often realized at home in family conversation or conversation among friends. Recordings included into ORAL2008 meet all these requirements and therefore ensure maximum possible authenticity of the language covered.

The second group includes sociolinguistic factors that influence character of spoken language on a more detailed scale. The most obvious is gender; many investigations focus on it nowadays. Another very important factor is age of speakers, as differences between generations indicate direction and form of language change. The third factor is education; it reflects social differences bound with a certain type of education. Finally, the fourth parameter are the regions of speakers' residence. The most influential of them is the region where they lived during the major part of their childhood, as it influences their idiolect furthest and for their whole life. The region where they lived in the period between their childhood and the time of recording affects their idiolect with less intensity and duration. Since place of birth itself can be accidental, it is considered the least important regional factor. Therefore, ORAL2008 will be fully balanced according to gender, age, education and region of childhood residence.

All other factors of the second group are considered of less relevance and are thus only registered in the database. These include occupation, number of speakers in transcription, particular type of situation and also additional technical data as length of recording etc. Some of them will be made available also to the query engine allowing their utilization for searching, making statistics etc.

4. Automatic balancing procedure of ORAL2008

As already mentioned above, all required information for every speaker is already recorded in the repository of source material for ORAL spoken corpora. However, the repository is unbalanced in terms of the required sociolinguistic factors (cf. above). The task is thus to select balanced subset of transcriptions of required overall size.

In order to keep both corpora disjoint, the transcriptions already used in ORAL2006 were excluded from the set of potential ORAL2008 candidates. Furthermore, some material not conforming to more specific conditions was excluded as well. The conditions were strictly informal recordings where all participating speakers are known to originate from some Bohemian region (speakers from Bohemian-Moravian transient region are not excluded, but their number is minimized, cf. below) with additional requirements concerning technical quality of the recording.

An automatic balancing program was run on the reduced set of candidate transcriptions trying to find its balanced subset sized one million running words proper. Balancing was required in the three basic binary sociolinguistic categories of speakers (gender, age and education), supplemented by the region of childhood residence (four Bohemian regions). An ideal subset of the candidate transcriptions should thus have 50% proportion of each value in the binary category and 25% proportion in the quaternary region category.

Computationally, this leads to the subset sum problem that is NP-hard, i.e. it is widely believed that the best solution can be found only in exponential time, which is not feasible given the size of the

transcription repository. Since the candidate transcription set is far from being balanced, random selection of transcriptions is also not possible and some kind of heuristics has to be employed instead. Therefore, the balancing algorithm was designed to be simple and effective, although it cannot be guaranteed to find the best solution.

The algorithm takes as its input list of all available transcriptions with respective word counts and sociolinguistic categories for every speaker in the transcription. Intuitively, since higher-educated younger women prevail in the transcription repository as a whole, transcriptions of lower-educated older men should be given priority for inclusion into the corpus. This is formalized by assigning a global «desirability coefficient» (DC) to each possible value of every individual sociolinguistic category: the rarer the category, the higher the DC and vice versa (e.g. older speaker +2, younger speaker –3, lower education +5, higher education –3 etc.). Overall «desirability coefficient» of the whole transcription (TDC) is naturally a compromise among the DC's of all individual categories of every participating speaker weighted by the number of words each of them utters. Provided the weighting scheme is invariable, a fixed set of global DC's determines value of TDC for every transcription and thus also ordering of the transcriptions according to the TDC values. The transcriptions can be then selected into the corpus starting from the one with the highest TDC until the required size of the corpus is reached. Therefore, every set of DC values in fact determines one particular subset of transcriptions.

The task for the balancing algorithm can thus be reformulated as finding the DC set that gives the best subset in terms of the required sociolinguistic categories. The algorithm runs in several passes trying to find the best DC set for each pass. The candidate DC sets are searched for in the vicinity of the temporary optimum from the previous pass (initial DC values are all set to zero). The algorithm iterates as long as the results improve, converging to a local optimum (that may not be global, though). The usual number of passes is 5–8, results of the balancing algorithm run on the current transcription

repository are shown in the following table. Although the results are not final yet as the transcriptions are currently being revised and slightly corrected, the final figures are expected to be very close.

Table 4. Number of words proper in selected categories in ORAL2008

Situation	**Formal**	**Informal**
	0	1 000 530
Gender	**Women**	**Men**
	498 000	502 530
Age	**Younger (18–35)**	**Older (36 and more)**
	500 181	500 349
Education	**Elementary/intermediate**	**University**
	499 954	500 576

Table 5. Number of words proper in region of childhood residence category in ORAL2008

Region	**Number of words**
Southwest Bohemia	244 584
Czech borderland	250 695
Northeast Bohemia	247 751
Central Bohemia	243 202
Bohemian-Moravian transient region	14 298

5. Conclusion

Corpus ORAL2008 was presented with its most distinctive features that draw together the advantages of its predecessors. It will contain only authentic spoken language used in informal situations recorded in the whole of Bohemia. It will be fully balanced according to gender, age, education and region of childhood residence of the speakers. Furthermore, the transcriptions will be aligned with corresponding recordings at word-level, thus making ORAL2008 an unprecedented source of information about spoken Czech language.