M. Kopřivová, M. Waclawičová

REPRESENTATIVENESS OF SPOKEN CORPORA ON THE EXAMPLE OF THE NEW SPOKEN CORPORA OF THE CZECH LANGUAGE[1]

1. Prague Spoken Corpus and Brno Spoken Corpus

Currently there are two corpora of spoken Czech in the Czech National Corpus[2] available, namely Prague Spoken Corpus built up in 1988–1996 and Brno Spoken Corpus built up in 1994–1999.

Prague Spoken Corpus[3] (PSC) is the first corpus of spoken Czech. It captures the authentic spoken language of Prague and its surroundings, which includes general language and is thematically unspecified. With reference to the special and geographically central position of Prague within the country, there are among others two significant points to be mentioned – inhabitants come from all parts of the Czech Republic and the city has an important medial influence on the rest of the country. Therefore, thanks to this variety of language, the corpus has to a large extent a nationwide character. For this corpus, 304 full anonymous recordings were made and subsequently transcribed. Given that these recordings were made between the years 1988 and 1996, they reflect the speech of the end of one social period and the beginning of another. Approximately half of them are thematically unspecified, capturing informal situations, half of them record formal dialogs based on questions that have been the same for each recording.

PSC was made with the aim of getting a balanced proportion of four main sociolinguistic parameters – speaker's gender, age, educa-

[1] This resource has been supported by the grant MSM 21620823.

[2] *Čermák F.* Today´s Corpus Linguistics. Some Open Questions // International Journal of Corpus Linguistics. 2003. Vol. 7. № 2. P. 265–282.

[3] *Čermák F.* Pražský muvený korpus. 2001. URL: **http://ucnk.ff.cuni.cz**

tion and type of discourse. All parameters were subdivided in a binary fashion. Each one of these four parameters can be shown within the search results using the Bonito corpus manager. Gender is represented with abbreviations M – Z (male – female). Age is tagged by I – V (younger – older). The category «younger» contains people from 20 to 35 years of age. This most limit was set taking into account that the speech of teenagers is not fully stabilized yet. The category «older» includes population over 35 years of age. Educational attainment is marked by B – A, where B stands for primary and secondary school and A for university. The last parameter is given by abbreviations F – N, meaning formal and informal speech. Formal speeches were monologs composed as answers to a series of questions. Questions were related to such topics as school, youth, employment, family, etc.

The number of all positions in PSC is 819 267 (i.e. number of all word forms and punctuation marks), while it includes 674 992 words.

Brno Spoken Corpus[1] (BSC) illustrates the authentic speech of the biggest city of Moravia, Brno. It is built up of electronic transcription of 250 anonymous recordings from the years 1994–1999, featuring 294 speakers. This corpus reflects the special position of Brno as a city in a dialectally rich region. There is a mixed dialect of central Moravia to be found here, and also common Czech and traces of an old slang of Brno (so called hantec) in the speech.

The corpus was built up so that the main four sociolinguistic parameters, adopted from PSC, were balanced. 135 recordings are formal and 115 informal.

A functional combination of phonetic transcription and standard spelling norms was used as the transcription method for both corpuses. The distinction between PSC and BSC consist of replacing

[1] *Hladká Z.* Tvorba a využití korpusů češtiny na FF MU v Brně // *Hladká Z., Karlík P.* (eds.). Čeština – univerzália a specifika 4. Praha, 2002. P. 307–310; *Hladká Z.* Brněnský muvený korpus. 2001. URL: **http://ucnk.ff.cuni.cz**

traditional punctuation with «pause» punctuation and recording simultaneity of dialogs. Complicated manual tagging and coding of corpuses is not finished yet, therefore it is possible to work with them only as plain texts, i.e. above all to search concordances with certain word form or their combinations.

2. Gathering data for the Czech Spoken Corpus

Since the year 2001 a series of recordings for a new corpus of spoken Czech language has been started. This time the aim is to acquire recordings from all parts of Bohemia (see Fig. 1). It will consist entirely of informal language, in contrast to PSC and BSC. The dialogues are mainly spoken in everyday unofficial and informal situations and their topics are not given beforehand. Emphasis is laid on the situations where the speakers know each other well and communicate in private. Formal situations are those, where at least one speaker acts in the terms of his or her profession (e.g. conversation with a doctor, hair-dresser, etc.). The selection of concrete situations is left to the recording persons. Generally it is not before the recording is finished, that the speakers are informed about the fact that their conversation was recorded and for what purpose. If the speakers express assent with their recording being used in the corpus, the recording is transcribed and inserted in the database of recordings.

Figure 1. Areas of recordings

Students are helping us with data acquisition; they are making records mainly in their domicile in various parts of Bohemia. Universities in five other cities in Bohemia participate on this project since the year 2005. This helps with a better representation of particular regions. Nevertheless, most recordings have been made in Prague. To a smaller extent (but not small) recordings represent regional variants of common speech. In the beginning, the recordings were made on cassettes, which were later digitized. Presently we use minidisks and MP3 recorders.

The transcription used in the corpus tries to be as authentic and comprehensible as possible. It originates from folkloristic transcription adjusted to the purposes of computer processing according to the usage established in the Czech National Corpus. It doesn't record intonation or other phonetic features as variations in pronunciation of phonemes.

Speakers are tagged with numerical codes, zero is reserved for the person conducting the recording. It could happen that this person, knowing that the dialog is recorded, would speak unnaturally or ask questions only to keep the conversation going. In such a case his or her utterance would have to be left out of transcription or signed as formal. It turned out that these cases barely appear in our recordings.

The actual transcription approximates traditional script, the main differences being the following: in the beginning of a sentence small letter is used with respect to computer processing. Capital letters are only used for proper names and some abbreviations. Sentences, which are interrupted or not finished (both cases are very frequent), are marked in a special way, with the help of graphical marks. There are cases, where the literary spoken form of language as well as the common spoken form differ from the script (*i – y*, *dě*, *tě*, *ně*, *bě*, *pě*, *mě*, *vě*, voiced and unvoiced sounds in certain positions). These we transcribe as in traditional scripts. But in other cases, where common speech regularly differs from literary pronunciation, we differ from the traditional script and transcribe differences as they occur in the

common pronunciation. Equally, we try to write down regional and dialectal features as well. Consequently doublets can appear in the transcripts. Thus we write *jsem* or *sem*, *půjdu* or *pudu*, *dole* or *dóle*, *zrovna* or *zrouna*.

The segmentation of continuous speech to graphic units depends largely on transcriber's interpretation. He or she makes decisions on the account of intonation, significant and graphical units. That means that speech is not segmented according to pauses, but in this regard it comes closer to the character of written texts.

Proper names are encoded because of necessity to safeguard anonymity of speakers and those, who are spoken about. Abbreviation NP is used instead of surnames, NN instead of nicknames. If the speakers don't wish their first names or geographical names to be made public, we use NJ instead of first names and NM in place of geographical names. Names of personalities and celebrities are not encoded.

Apart from technical specifications of the recording (length of the recording, month and year of recording, place and area) we observe the language situation where we indicate the type (formal or informal), topic, physical presence of the speakers (which may be a telephone conversation) and preparedness of the speaker. Further we note whether it is a dialog (2 speakers) or more speakers are present, the relationship between the speakers (if they don't know each other, they do know each other, they are friends) and the environment (private or public). The most common situation is a conversation during a meal at home or in a restaurant.

Further on, anonymous information about the speakers is mentioned in the database. Analogous to PSC and BSC, it is their gender (male or female), age, education (elementary school, high school, university), place and region of birth, place of residence during childhood. For the region of birth we use Bělič's division of

regions – Central Bohemia, Northeast Bohemia, Southwest Bohemia, Central Moravia, East Moravia and Silesia and the border areas[1].

3. The Czech Spoken Corpus

The first part of the spoken corpus ORAL2006[2] will be made accessible by the end of 2006 under the corpus manager Bonito in the same way as the other corpuses of the Czech National Corpus (CNK). Similarly to PSC and BSC, it involves transcriptions of about 250 recordings from informal situations.

The corpus ORAL2006 is created from material that is rather unbalanced in the typology of situations as well as speakers. The majority of recordings come from family conversations during a meal at home or from conversations in a restaurant or at a college.

It would be more appropriate to use sociolinguistic categories of the speakers (gender, age, education) as a basis for creating a new corpus. It would allow for comparisons with existing corpora of spoken Czech language. However, the recordings are unbalanced in this aspect as well. College students record conversations with their peers or with their family members, and so speakers under the age of 35 with university education predominate in our material. Currently we are trying to recruit also older colleagues for recordings, but it is not easy to train them to master the methodology of gathering and transcription of the recordings.

The category of the origin of the speaker according to dialectal areas was incorporated into the criteria for corpus creation, too. At first, recordings have been conducted in Prague and as late as since 2005 in other cities too, therefore the corpus will contain predominantly transcriptions of recordings from Central Bohemia.

---

[1] *Bělič J.* Nástin české dialektologie. Praha, 1972.

[2] *Čermák F., Sgall P.* Výzkum mluvené češtiny: jeho situace a problémy // SaS. 1997. № 58. P. 15–25.

Nonetheless, smaller quantity of recordings from other Czech dialectal areas will be included as well.

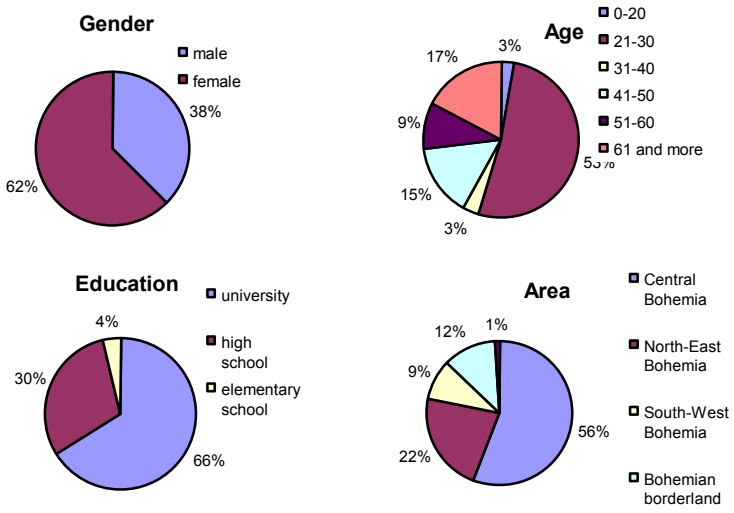

Figure 2. Structure of ORAL2006

There are 98% dialogs that are informal and 2% formal dialogs in a bank of spoken corpus recordings, which have been made till this time. 96% dialogs are among people who are friends, in 2% dialogs they know each other and in 2% that don't know each other at all. There are 42% male and 58% female speakers, 72% speakers are younger than 35 years of age, 20% are between 35 and 60, 8% are over 60. 69% speakers have been studying university, 26% has secondary education and 5% primary education. 58% speakers come from Central Bohemia, 17% from North-East Bohemia, 16% from Bohemian borderland, 7% from South-West Bohemia, 2% from Moravia.

The corpus ORAL2006 will be composed according to the typology of speakers[1]. We have chosen about 80% of all dialogs and have reached bigger proportion of representativeness. This selection in this stage is not ideally representative yet and we are working on improving it. The current preliminary form is shown in Fig. 2.

4. Conclusion

The new corpus ORAL2006 will be useful for a broad spectrum of research. The collected data under the corpus manager Bonito will show us the character of lexicon, morphology and syntax in spoken language. It will be possible to make frequency analysis and compare it with corpora of written language. Thanks to the information about gender, age, education and place of origin we will be able to examine differences depending on these sociological parameters. In regard to repeating situations it will be conceivable to research phraseology and typical conversation in these everyday situations. The nivelisation of language in Prague and border areas will be well shown in combination with the influence of the region, from where the speakers come from, of the firmness of particular regional features and of the neighboring regions. Comparison with PSC and BSC that consist mostly of urban speech, originate from another period and contain mostly formal speech will be publicly accessible.

[1] The Corpus of Spoken Israeli Hebrew (URL: **http://www.tau.ac.il/humanities/semitic/cosih.html**) uses the typology of speakers as one of the criteria of representativeness.