Balanced corpus of informal spoken Czech: compilation, design and findings

Martina Waclawičová, Michal Křen, Lucie Válková

Institute of the Czech National Corpus, Charles University in Prague, Czech Republic
{martina.waclawiczova,michal.kren,lucie.valkova}@ff.cuni.cz

Abstract

The paper presents ORAL2008, a new 1-million corpus of spoken Czech compiled within the framework of the Czech National Corpus project. ORAL2008 is designed as a representation of authentic spoken language used in informal situations and it is balanced in the main sociolinguistic categories of speakers. The paper concentrates also on the data collection, its broad coverage and the transcription system that registers variability of spoken Czech. Possible findings based on the provided data are finally outlined.

Index Terms: spoken language, transcription, corpus design, sociolinguistic representativeness

1. Introduction

Growing interest in spoken language as the primary means used in everyday communication is further supported by its importance for linguistic research as well as natural language processing in general [1]. Paradoxically, resources of exactly this kind of language are scarce as they are very expensive to build. The Czech National Corpus (CNC) project aims to provide resources for description of wide varieties of Czech language including also authentic spoken language used in informal situations. The spoken data collection project is already established as a continuous mapping of contemporary spoken language on a regular basis.

Spoken language research is particularly important in the Czech language situation, for it is sometimes described as close to diglossia [2,3]. Spoken Czech language is typically non-literal and shows huge variability – phonetic, morphological, syntactical and lexical. Moreover, it is not only simply the distinction from the written language; spoken Czech shows also variability within the system. It is mostly bound with regional speech differences as a result of former dialectal diversity. Morphological system is probably the most remarkably different from the written one. Last but not least, the syntax of spoken varieties of every language has different general rules, because it is transferred through the sound channel and the speech is both produced and perceived differently than written language.

The first part of this paper concentrates on design fundamentals of ORAL2008, a new sociolinguistically balanced corpus of informal spoken Czech covering the whole of Bohemia and sized 1 million words. The second part examines the nature of provided corpus data and outlines also possible linguistic and sociolinguistic findings based on them.

2. Background

Apart from ORAL2008, the following spoken corpora are available for research purposes at the CNC:

- Prague Spoken Corpus (PSC) [4] was recorded in Prague in 1988 – 1996 and contains 819 267 tokens, of which 674 992 are words proper (tokens that contain at least one alphabetic character, i.e. the punctuation is excluded). It is built of transcriptions of recordings of both informal and formal situations, the latter being a controlled dialog with topic given in advance. Simple annotation of the corpus gives information about (in)formality of situation and basic binary sociolinguistic categories of the speakers: gender, age group (older / younger) and education (lower / higher). The corpus is relatively balanced in these four parameters.

- Brno Spoken Corpus (BSO) [5,6] was recorded in Brno in 1994 – 1999 and contains 596 009 tokens (489 410 words proper). Its design stems from PSC so that they are comparable in many respects, including also both formal and informal parts.

- ORAL2006 [7,8] was recorded in the whole of Bohemia in 2002 – 2006 and it contains 1 312 282 tokens (1 000 798 words proper). It was the first of the ORAL-series corpora that resulted from the CNC spoken data collection project. Therefore, it diverges from the original PSC design, mainly that it consists entirely of informal spoken Czech. Unfortunately, ORAL2006 is not balanced due to insufficient material available at the time.

3. Corpus ORAL2008

ORAL2008 [9] is the second ORAL-series corpus of informal spoken Czech sized 1 000 097 words proper that has been made available in December 2008. ORAL2008 was 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 are compatible in all respects, including the transcription rules and overall annotation scheme. However, the individual recordings and transcriptions already included into ORAL2006 were not re-used in ORAL2008, so that there is no intersection between the two corpora. Moreover, ORAL2008 is fully balanced in the main sociolinguistic categories of speakers – gender, age group, education and region of childhood residence (cf. below).

3.1. Data collection and variability

The spoken data collection project coordinated by the CNC aims at collecting material that would represent authentic spoken language used in informal situations. The necessary prerequisite is having recorded and transcribed a large variety of speakers from various regions in the repository [10]. For this purpose, a network of local collaborators was established and the repository size is still increasing. The recordings and transcriptions are collected also at five regional universities, the total number of people involved in making the recordings and transcriptions is about 170. Such a wide-scale project requires not only a great deal of organization and administration overhead, but also an appropriate computer support including a central data storage and standardization tools.

Distributed character of the project is supported by the database system used for storing the recording, its
transcription and also additional information about the situation, participating speakers etc. The database also asks for confirmation of an on-line statement that all the speakers agreed with inclusion of the recording and the transcription into the CNC. This is required before a new record can be created as an addition to the signed written statement received by regular mail.

The database system is a set of PHP forms and additional Perl scripts above MySQL database accessible online via https protocol [11]. All the submissions are made through the database that ensures specified format of the sound files (uncompressed 16-bit PCM WAV, mono, 16 kHz) and formal conformance of the transcription files. The formal conformance includes not only XML parsing, but also basically every automatically-checkable feature laid out in the transcription manual. Furthermore, all the transcriptions are made and time-aligned manually using Transcriber [12], and subsequently verified by the local coordinator as well as the central administrator before they are finally approved, so that maximum possible reliability of the transcriptions is ensured.

The additional information stored in the database can be divided into three groups. The first group comprises technical data about the recording, such as its length, as well as the month, year, place and region of the recording.

The second group focuses on the language situation and the following items are observed: general type of situation (formal / informal, private / public), particular type of situation (visit, restaurant, celebration, trip etc.), topic of dialogue, physical presence of speakers and preparedness of speech. Number of participating speakers and the relationship among them is also noted.

The third group comprises information about the speakers, namely their gender, exact age, age group (younger (18–34) or older (35 and more)), education (elementary school, high school or university), type of current occupation, place and region of birth, regions of childhood and current residence. All other factors are selected as balancing criteria, while the other ones usually vary arbitrarily.

Apart from the basic functions such as storing and searching the data, the database also registers counts of words uttered by the individual speakers in each transcription file. This information is not only displayed, but also linked with the sociolinguistic categories of the speakers and provided to the central administrator as well as the local coordinators that can thereby continually balance the data composition. Since there is no personal information stored in the database, a semi-automatic detection of duplicate speakers is applied. The word counts are then used for subsequent restrictions on the total number of words each individual person utters in the collected material.

3.2. Representation of authentic spoken language

As mentioned above, ORAL2008 is designed to represent authentic and prototypical spoken Czech used in informal situations [15]. Character of spoken language is strongly influenced by a variety of factors [16] that can be divided into two groups in relation to the needs of corpus compilation. Factors of the first group should be observed to ensure the prototypical spoken language as opposed to the written language and thus constitute suitable selection criteria. Factors of the second group influence character of the spoken language on a more detailed scale and can be used for corpus balancing. In practice, however, only the most important factors are selected as balancing criteria, while the other ones usually vary arbitrarily.

3.2.1. Selection criteria

The first group consists of the following factors that characterize the prototypical spoken language. The key factor all the other ones are related to is informality of situation. Formal situations [17] require language of almost written character, while the authentic spoken language is bound with informal situations. The other factors include private environment and unscripted speech, as well as topic not given in advance and physical presence of speakers. Relationship between the speakers is also affected, as they know each other well in such a situation. All these factors are mostly bound with dialogical, not monological character of speech. In practice, they are most often realized in talk within a family or among friends. Recordings included into the ORAL corpora meet all these requirements that ensure maximum possible authenticity of the language covered.

3.2.2. Balancing criteria

Factors of the second group include most obviously gender, age, education and region of residence. Differences between generations naturally indicate direction and form of language change, while education reflects social differences bound with a certain type of education. For the sake of simplicity, plain PSC-based binary values of age and education (cf. Section 2) were used for balancing purposes. As for the region of residence, the most influential is the region where speakers spent 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 the recording affects their idiolect with less intensity and duration. Since place and region of birth can be accidental, they are considered the least important regional factors.

Therefore, ORAL2008 is fully balanced in gender, age group, education and region of childhood residence. All other factors, such as occupation, particular type of situation, number of speakers etc., are considered of less relevance and are thus only recorded in the database.

3.3. Automatic balancing procedure

The repository of the material collected for the ORAL corpora is not balanced in terms of the required sociolinguistic
categories. Because of the balancing requirements for ORAL2008, it was necessary to select a subset of the transcriptions with equal (or almost equal) proportion of words for each value within every relevant sociolinguistic category (e.g. 50% of words in the corpus uttered by men, 50% by women, 25% for each of the four childhood residence areas etc.). Naturally, every transcription is indivisible, it can be selected only as a whole, so the balancing can hardly be done manually. Such a task computationally leads to the subset sum problem that is NP-hard, i.e. the best solution can be found only in exponential time. Since this is not feasible given the transcription repository size, a heuristic method was applied instead. The algorithm [18] was designed to be simple and effective, although it cannot be guaranteed to find the best solution. However, the results turned out to be very good and are shown in the following tables.

Table 1. Number of words proper in selected binary categories in ORAL2008.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>500 478</td>
<td>499 619</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age group</th>
<th>Younger</th>
<th>Older</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>500 199</td>
<td>499 898</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education</th>
<th>Lower</th>
<th>Higher</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>499 425</td>
<td>500 672</td>
</tr>
</tbody>
</table>

Table 2. Number of words proper in region of childhood residence category in ORAL2008.

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Bohemia</td>
<td>247 630</td>
</tr>
<tr>
<td>Northeast Bohemia</td>
<td>242 199</td>
</tr>
<tr>
<td>Southwest Bohemia</td>
<td>237 417</td>
</tr>
<tr>
<td>Bohemian borderland</td>
<td>239 511</td>
</tr>
<tr>
<td>Bohemian-Moravian</td>
<td>33 340</td>
</tr>
<tr>
<td>transient region</td>
<td></td>
</tr>
</tbody>
</table>

The adopted approach with equal sizes for each value in every sociolinguistic category may be questionable from sociological point of view. However, any numbers regarding “ideal” representation in terms of e.g. education could be disputed as well, while the same-sized portions are at least very practical for end users offering them an easy way to compare various frequency characteristics in relation to these categories.

4. Transcription rules

Characteristic features of spoken language are primarily available by means of special transcription in spoken corpora [19]. Given the broad coverage of the project and the number of people involved, the transcription system should be not only well-grounded linguistically, but also easy to learn for individual transcribers and comprehensible for corpus users. In other words, it is important to realize the trade-off between required amount of collected material on one side and depth and sophistication of the annotation on the other. Therefore, it should be stressed that it was not aimed to build a multi-modal corpus with several layers of detailed annotation for studies of phonetics or pragmatics. The material is rather primarily intended to facilitate studies of morphology, syntax, lexicon and syntagmatics of spoken language in contrast with written language. For this purpose, corpus size and variability are of fundamental importance.

There is only one transcription layer in ORAL-series corpora that is manually transcribed and verified. The transcription is based on the Czech traditional script that is often etymologic or morphologic rather than phonetic. The transcription aims to capture variability of spoken language in contradiction to the traditional script. However, the variability is relatively high and it can easily merge with cases of indistinct or vague pronunciation. Therefore, a set of rules was elaborated to delimit the boundaries of variation that is to be recorded in the transcription. Furthermore, the transcribers are provided with a list of the most frequent and the most typical cases of different types of pronunciation and their transcription, so that the room for their own decisions based on the general rules is limited.

Words that are pronounced the same way in codified literary form of language as in common spoken form are written according to the traditional script. Cases where the literary spoken forms as well as the common spoken forms differ from the script (e.g. i – y, dě, tě, ně, bě, pé, mé, vě, voiced and unvoiced sounds in certain positions) are transcribed in traditional script as well. Nevertheless, all the other cases where common speech regularly differs from literary pronunciation are transcribed with all differences as they occur in the common pronunciation. This holds equally for regional or dialectal distinctions. Hence all differences in simplifications of consonant groups are registered, e.g. jsen vs. sem (“I am”), pijdu vs. pudu (“I will go”), as well as differences in vowel length, e.g. dole vs. důle (“down”), prosim vs. prosim (“please”), in assimilation of consonants, e.g. s mášem vs. z mášem (“with butter”) or regional pronunciation zrova vs. zrova (“just”) etc. As a consequence, many doubles appear in the transcripts that reflect the actual pronunciation.

Non-proper names are always written in lowercase. Interrupted or unfinished utterances (very frequent in common speech) are marked with special punctuation symbols. Segmentation of continuous speech to graphic units depends largely on transcribers’ interpretation. They make decisions on the account of intonation, syntactical units and meaning. That means that speech is not segmented according to pauses, but it comes closer to the character of written texts in this respect.

5. Findings based on the provided data

All the corpora at the Czech National Corpus are available through Bonito, a graphical user interface to corpus query engine Manatee [20]. For the time being, its major drawback regarding spoken corpora is lack of functionality to utilize alignment of the transcriptions with the sound files that would enable to hear the actual realizations. On the other hand, it is a powerful tool that supports complex regular expression queries, various statistical functions including frequency distribution or collocation measures, creating of subcorpora using external annotation etc. This enables end users to easily obtain information about e.g. speech variants and their frequency distribution according to the sociolinguistic categories of speakers, as well as research of regional and dialectal features. For instance, according to the studies of the oldest traceable stage of the dialects, dative suffix of masculine animate substantives -oj is a dialectal feature typical for the Northeast part of Bohemia. Consulting a corpus, it can be assured that this feature is still productive nowadays being used in 15% of all datives in the whole Bohemia. Further, it is still characteristic for its original region, for 91% of its occurrences come from the Northeast Bohemia, 5% from the
Central Bohemia and only marginal occurrence is registered elsewhere.

As mentioned above, additional information about the situation and participating speakers is recorded during the data collection. Most of this annotation is made available also to the query engine for sociolinguistically oriented research of frequency characteristics of speech. It includes the year of the recording and the number of speakers involved, as well as their gender, exact age, age group, type of education and region of their childhood residence. This enables for instance creating subcorpora of utterances of particular type of speakers and easy obtaining relevant frequency information for research on lexical, morphological or phonetic features that embody some kind of variability in relation to the sociolinguistic characteristics.

For instance, it can be attested that women tend to use more ensuring and contact words – věd' ("am I right?" or "doesn’t it?") is used by women in 63% of all cases. Investigating influence of the age of speakers on lexical level can assure that usage of “trendy” words prevails in young generation utterances, e.g. hustej “cool” is used in 78% of all cases by speakers in younger age group.

Using a combination of age and region it can be attested that some dialectal features (mainly those most regionally restricted or noticeably different from literary Czech) are characteristic for speech of older generations and disappear from younger people discourses. For instance, dolu “downwards” has a dialectal variant with vocal length dolu, and this variant is in 82% used by older people and in 82% by people grown up in the Northeast Bohemia and in the Bohemian borderland. Finally, a detailed research was carried out that aimed at speech in the Bohemian borderland where no dialectal subsurface exists. Occurrence of different variants observed in different parts of the area pointed to a quite new fact: the Bohemian borderland divides into two parts, each of them being strongly influenced by the neighboring dialectal region [21]. This finding is based entirely on the corpus data and it is convincing due to their size, coverage and overall reliability.

6. Conclusion
Corpus ORAL2008 was presented as an indispensable source of data representing contemporary authentic spoken Czech. This is achieved by the broad regional coverage, detailed annotation, large variety of recorded speakers and final balancing of the collected material. The amount of data is growing steadily for a new corpus that will join ORAL2006 and ORAL2008 in the future.

Currently, preliminary works are being done to adapt tools used for lemmatization and morphological tagging of written Czech [22,23] to spoken corpora. So far, it is not possible to use their unmodified versions because of poor coverage of spoken word forms and different character of spoken language in general. After this will be at least partially resolved, a lemmatized and morphologically tagged super-corpus ORAL will be published. It is planned to be a unification of all ORAL-series corpora, regularly updated and re-processed with the newest versions of available tools.

7. Acknowledgements
This research has been supported by MSM0021620823 grant.

8. References