Balanced data repository of spontaneous spoken Czech

Lucie Válková, Martina Waclawičová, Michal Křen
Institute of the Czech National Corpus, Charles University in Prague
Nám. Jana Palacha 2, 116 38 Praha 1, Czech Republic
{lucie.valkova,martina.waclawicova,michal.kren}@ff.cuni.cz

Abstract

The paper presents data repository that will be used as a source of data for ORAL2013, a new corpus of spontaneous spoken Czech. The corpus is planned to be published in 2013 within the framework of the Czech National Corpus and it will contain both the audio recordings and their transcriptions manually aligned with time stamps. The corpus will be designed as a representation of contemporary spontaneous spoken language used in informal, real-life situations on the area of the whole Czech Republic and thus balanced in the main sociolinguistic categories of speakers.
Therefore, the data repository features broad regional coverage with large variety of speakers, as well as precise and uniform processing. The repository is already built, basically balanced and sized 3 million words proper (i.e. tokens not including punctuation). Before the publication, another set of overall consistency checks will be carried out, as well as final selection of the transcriptions to be included into ORAL2013 as the final product.

Keywords: spontaneous spoken language; Czech; balanced corpus data

1. Introduction

Spontaneous spoken language is the primary means of everyday communication, and this fact leads to growing, many-sided research interest in its linguistic descriptions, as well as to their applications in natural language processing. The importance of spoken language data is further supported by the significant differences between informal spoken language on one side and rather formal written language on the other. These differences are primarily caused by the fact that the spoken language is transferred through a sound channel and the speech is thus produced and perceived differently than writing. As opposed to the written language varieties, spoken varieties of every language show a number of differences mainly in syntax, or rather patterns of usage (Miller & Weinert 1998). This holds especially for spoken language used in informal situations that exhibits a number of specific features only rarely present in written language (e.g. heavy implicit reliance on context, fragmented and unintegrated utterances, hesitations, rectifications, smaller range of vocabulary etc.). This difference means, among other things, that spoken corpora are indispensable source of information about spontaneous spoken language that can hardly be derived from written language data or otherwise inferred from other sources.
Specificity of the Czech language situation should be pointed out in this context. Apart from the dialects, there are two central language formations in Czech, the (mostly written) literary standard and the (mostly spoken) colloquial language called Common Czech (Sgall et al. 1992). There is a substantial gap between them and this situation is sometimes described as close to diglossia (Bermel 2010). The differences between spoken and written Czech are in general two-fold. First, they are caused by the channel of communication (which is common in any language) and their nature could be (in a simplified way) described as syntactic. Second, colloquial Czech is specific also on phonetic, lexical, and mainly morphological level. As Czech is a highly inflected language, this makes the two formations very distinct, and there are often expressions that are either marked as literary or colloquial. For instance, "děkuji" / "děkuju" are formal / informal variants for "thank you" distinguished by different suffix and – most importantly – there is no neutral, non-marked form. It is also typical for a number of very common concepts to be expressed by different lexemes with often complementary distribution in spoken vs. written language, e.g. "stále" / "pořád" / "furt" for "all the time, still". Moreover, the differences should not be regarded as binary (typical either for spoken or written language), but rather as a more complex set of relations, especially when considering also the gradually disappearing, but still present regional varieties. This underlines the importance of the regional factor, in addition to other sociolinguistic variables, as well as their balance, for designing a data repository that would aim to represent spontaneous spoken Czech properly and in its entirety.
Although the notions of representativeness and balance in corpus design are often questionable as they cannot be objectively measured, we claim that they are also unavoidable at the same time. They become a key issue whenever it is necessary for research findings, n-grams or language models based on the corpus as a language sample to be applied to the language as a whole. This leads to the need of representative and balanced corpora as a valuable source of language data for both basic and applied research. Not unexpectedly, resources of this kind are very scarce especially for spontaneous spoken language, as their compilation is very laborious and thus expensive. The presented data repository aims to bridge this gap as a continuation and enhancement of the corpus-building activities described in the next section.

2. Background

One of the main aims of the Institute of the Czech National Corpus (ICNC) is continuous mapping of contemporary language. This effort results in compilation, maintenance and providing access to a number of corpora that represent different language varieties. One of the varieties is spontaneous spoken language used in informal situations and it is covered with corpora that make up the ORAL series.

Until today, two corpora based on this data have already been published: ORAL2006 and ORAL2008. They are each sized 1 million words proper (i.e. tokens not including punctuation) and built from the material recorded in the whole of Bohemia (i.e. bohemian part of the Czech Republic, not Moravia and Silesia) using the same repository of recordings; throughout this paper, the word 'recording' will be used as a term referring to both audio and its manual transcription (Crowdy 1994). This means that the two corpora are compatible in all respects, including the transcription guidelines and overall annotation scheme. However, the individual recordings already included into ORAL2006 were not re-used in ORAL2008, so that there is no intersection between the two corpora.

ORAL2006 (Kopřivová & Waclawičová 2005; Waclawičová 2007) was published in November 2006, it consists of the material recorded 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 ICNC spoken data collection project.

ORAL2008 (Waclawičová & Křen 2008; Waclawičová, Křen & Válková 2009) was published in December 2008, it consists of the material recorded in 2002–2007 and it contains 1 349 536 tokens (1 000 097 words proper). Moreover, ORAL2008 is fully balanced in the main sociolinguistic categories of the participating speakers: gender, age group, education and region of childhood residence.

Both ORAL2006 and ORAL2008 are valuable and widely used resources that contribute to the description of spontaneous spoken Czech in general and also help to reveal facts not known so far, e.g. dialectological indication of two divergent tendencies in Bohemian borderland caused by the neighbouring regions (Waclawičová 2009). The findings are supported by adequate regional coverage, variety of recorded speakers and also balancing of the data in case of ORAL2008.

However, ORAL2006 and ORAL2008 also suffer from basically three major drawbacks: they do not cover the whole of the Czech Republic (but only Bohemia, i.e. about 2/3 of the area), the transcriptions are not aligned with audio and there is only one transcription layer that is close to the standard ortography. The reasons for this are manifold, but most importantly, data collection for the ORAL-series corpora started in 2002 in very modest conditions and until the end of 2007 it grew basically only quantitatively. The focus on the amount of the data, was caused by the primary intention to facilitate studies of variability and diversity of spoken language in contrast with written language. This orientation was revised by the end of 2007 and, as a result, building the original repository was discontinued at that time. One year later, ORAL2008 was published as the last corpus based on it.

At the beginning of 2008, several important changes were adopted. First, all the transcriptions started to be manually aligned with audio. Second, the regional coverage extended to the whole of the Czech Republic (thus including also Moravia and Silesia). Third, the inclusion of non-Bohemian regions caused some minor changes of transcription guidelines (taking into account regional phonetical differences) that developed into switchover from traditional to pause-based punctuation. Since then, all the data have been prepared according to the new standards.

The organization of the project and its overall nature were not changed, though. As it was already firmly established, the spoken data collection continued on a regular basis. The outcome of this effort is the new data repository that will be described in detail in the next section. It incorporates all the changes mentioned above and it is thus different from the original data repository used for compilation of ORAL2006 and ORAL2008.

3. Description of the data repository

3.1 Data collection and variability

The data repository for future ORAL-series corpora aims at collecting material that would represent spontaneous spoken language used in informal situations. The necessary prerequisite is having recorded and transcribed utterances of a large variety of speakers from various regions in the repository (Gibbon, Moore & Winski 1998). For this purpose, a network of local collaborators was established. The recordings are collected also at five regional universities, the total number of people involved in making the recordings is almost 200. 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 audio, its transcription and also additional information about the situation, participating speakers etc. (Křen & Waclawičová 2011) The database also asks for confirmation of an on-line statement that all the recorded speakers agreed with inclusion of the audio and its transcription into the ICNC. 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. All the submissions are made through the database that ensures specified format of the sound files 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 (Geoffrois et al. 2000), 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 communication 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, age, education (elementary school, high school or university), type of current occupation, place and region of birth, regions of childhood and current residence. The regions follow the traditional dialectological division – Central Bohemia, Northeast Bohemia, Southwest Bohemia, Bohemian borderland, Bohemian-Moravian transient region, Central Moravia, Eastern Moravia, Silesia and Moravian borderland (Bělič 1972; Balhar et al. 1992–2005).

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 thus continually balance the data composition. Since there is no personal information stored in the database, a semi-automatic supervised 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 Representativeness and balance

The data repository is designed as a representation of spontaneous spoken Czech used in informal situations, as this language variety is considered prototypically spoken as opposed to the written language (Čermák 2009). Obviously, character of spoken language is strongly influenced by a variety of factors. These factors can be divided into two groups in relation to the data repository construction needs.

Factors of the first group should be observed to ensure the prototypicality of spontaneous spoken language and thus constitute suitable selection criteria. The key factor is informality of situation (Labov 1972), 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 data repository meet all these requirements that ensure maximum possible authenticity of the language covered.

Factors of the second group influence character of the spoken language on a more detailed scale and can be used for balancing purposes. These include most obviously gender, age, education and region of residence (more precisely, it is the region of childhood residence as it influences the speaker's idiolect the most). As a practical solution, only these four factors were selected as balancing criteria, while the other ones can vary arbitrarily. For the sake of simplicity, plain binary values of age group (younger [18–34] / older [35 and more]) and education (lower / higher) were used for balancing purposes, in addition to the binary values of gender. It means that 50% representation of each of these values is considered the "ideal" balance (e.g. 50% of words in the repository uttered by men, 50% by women).

The 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. This approach was already adopted for ORAL2008, with the exception of the region of childhood residence that now includes also Moravia and Silesia and constitutes all 9 dialectological areas with sometimes significantly different extent and number of inhabitants.

As a result of these considerations, efforts were made to continually balance the data repository in terms of gender, age group, education and region of childhood residence during its compilation. All other factors, such as occupation, particular type of situation, number of speakers etc., were only registered as a part of the information about the speakers and the recording. The following tables show the balance of the data repository:

Gender	Women	Men
	1 465 528	1 558 547
Age group	Younger	Older
	1 618 333	1 405 742
Education	Lower	Higher
	1 622 454	1 401 621

Table 1: Number of words proper in selected binary categories in the data repository.

Region	Number of words
Central Bohemia	628 056
Northeast Bohemia	386 028
Southwest Bohemia	342 456
Bohemian borderland	208 669
Bohemian-Moravian transient region	82 937
Central Moravia	559 294
Eastern Moravia	377 084
Silesia	336 598
Moravian borderland	100 536

Table 2: Number of words proper in selected binary categories in the data repository.

3.3 The data

The data consists of 940 recordings with 2 877 speakers (out of which are 1 524 unique speakers) that include both audio and its manual transcription in Transcriber XML format. The total length of audio is 324 hours, the total size of the transcriptions is 3 024 075 words proper, i.e. about 4 million tokens including punctuation. All the audio files are in uncompressed 16-bit PCM WAV, mono, 16 kHz format. Although it was attempted to produce quality audio files, it was sometimes not possible to avoid background noise that is necessarily present in the real-life situations.

The transcriptions are manually aligned with audio on the level of segments. Segment is a sequence of typically 5–10 words that constitute a natural unit, be it prosodic, syntagmatic or semantic. Maximum length of the segment is limited to 15 words and the segment boundaries are always within a single speaker's turn.

There is only one transcription layer that is manually transcribed and double-checked. Roughly speaking, the transcription system is based on the standard Czech ortography and aims to capture the speech in writing, so that it would enable reconstruction of the original pronunciation. The standard ortography is thus not adhered to in case of non-standard pronunciation (e.g. consonant group simplifications, vowel length differences, assimilation of consonants etc.) or cases when pronunciation commonly varies (e.g. pronunciation of abbreviations) – for more details, please refer to Waclawičová, Křen & Válková (2009). For this purpose, a set of rules was elaborated in order to delimit the boundaries of variation that are to be registered in the transcription consistently. The transcribers were provided with a list of the most frequent and the most typical cases of different types of actual pronunciation and its transcription to limit the room for their own decisions based on the general rules.

It should be noted that the transcription system does not exclude further automatic processing of the data. The transcription can also be characterized as close to the ortography used typically in fiction to underline the colloquiality of the language used. As it is desirable to be able to process also this kind of language, the existing lemmatization and POS-tagging tools – and in particular, the morphological analysis module (Hajič 2004; Jelínek 2008) – will be further improved to cover the resulting ortographic variability. This has already been done for frequent word forms during preparation of lemmatized and POS-tagged spoken language data used as a base for the recent Frequency Dictionary of Czech (Čermák, Křen et al. 2011).

The punctuation is pause-based, distinguishing shorter and longer pause relatively to the individual pace of speech. Interrupted or unfinished utterances, which are very common in spontaneous speech, are also marked with special punctuation symbols.

Proper names are often anonymized in the transcriptions. As a rule, surnames are always anonymized, while anonymization of first names, nicknames, names of places etc. is optional and depends on consideration of the person who makes the transcription, as well as on possible requests from recorded speakers. The anonymization markers always occur in a special segment to facilitate automatic substitution of the corresponding part of the audio file.

Generally, the transcription system is designed to be not only well-grounded linguistically, but also easy enough to learn for individual transcribers and comprehensible for corpus users. It is a practical solution given the broad coverage of the project, the number of people involved, as well as the overall project priorities. Therefore, it should be stressed that it was not aimed to build a data repository that would be multi-modal with several layers of detailed annotation, but rather relatively large for studies of morphology, syntax, lexicon and syntagmatics of spoken language. For this purpose, data size and variability are of fundamental importance.

4. Availability

The data repository itself will not be published. However, it is planned to publish corpus ORAL2013 as the final product. It will be released in 2013 on http://www.korpus.cz/ as one of the corpora available to all registered users of the ICNC. The anonymized recordings (i.e. audio with the transcriptions) with all the (anonymous) information about the speakers will also be made available as a dataset suitable for NLP. The licensing is not decided yet, but we expect rather restrictive license for the dataset as a whole, and very permissive license in case of derived data, e.g. n-grams.

However, before ORAL2013 will be ready, quite a bit of work needs to be done. First and most importantly, the data will require another set of overall consistency checks necessary in a numerous cases where the language variability allows for multiple transcriptions that can be easily confused, yet there is no way how to check this automatically. For instance, "kdy byste" / "kdybyste" for "when would you" / "if you" can be distinguished only on the basis of meaning.

Second, although the data repository is basically balanced, the balance is not ideal and it is thus planned to select its subset as a future corpus ORAL2013. The ideal balance proportions are not determined yet, especially for the

region of childhood residence category. However, the overall size of ORAL2013 is expected to be close to the size of the data repository. Naturally, every transcription is indivisible and it can be selected only as a whole. We expect the algorithm used for balancing ORAL2008 to be adapted and applied on the data repository to select its subset for ORAL2013, as the balancing can hardly be done manually (Waclawičová & Křen 2008).

5. Conclusion

The presented data repository will be used as a source of data for a new corpus ORAL2013 that will be published in 2013. Because the corpus as the final product aims to represent spontaneous spoken Czech in a sociolinguistically balanced way, the data repository itself is already basically balanced in the main sociolinguistic categories of speakers. Furthermore, it features manual annotation, broad regional coverage and large variety of recorded speakers.

The repository is already built, balanced and sized 3 million words proper. However, it is not ready for publication yet. Work to be done includes mainly another set of overall consistency checks and final selection of the recordings (i.e. audio manually aligned with its transcription) that will be included into ORAL2013. We believe that the corpus will prove to be an indispensable source of data for various kinds of linguistic studies, as well as for many fields of natural language processing.

Apart from the compilation of ORAL2013, the future plan is to tackle the problem of missing phonetic information about the actual realization. As the project will hopefully continue for another five years, another manually-encoded layer will be added to the newly-made recordings to provide also this kind of information, although the details are still to be discussed.

6. Acknowledgements

Compilation of the data repository and its further development will be supported as a part of the Czech National Corpus project within the framework of Large Research, Development and Innovation Infrastructures.

7. References

Balhar, J. et al. (1992–2005). *Český jazykový atlas*. Praha: Academia.

Bělič, J. (1972). *Nástin české dialektologie*. Praha: SPN.

Bermel, N. (2010). O tzv. české diglosii v současném světě. *Slovo a slovesnost*, 71(1), pp. 5–30.

Crowdy, S. (1993). Spoken Corpus Design and Transcription. *Literary and Linguistic Computing*, 8(4), pp. 259–265.

Čermák, F. (2009). Spoken Corpora Design. Their Constitutive Parameters. *International Journal of Corpus Linguistics*, 14(1), pp. 113–123.

Čermák, F., Křen, M. et al. (2011). *A Frequency Dictionary of Czech: Core Vocabulary for Learners*. London: Routledge.

Geoffrois, E., Barras, C., Bird, S., Wu, Z. (2000). Transcribing with Annotation Graphs. In *Proceedings from The Second International Conference on Language Resources and Evaluation (LREC)*, pp. 1517–1521.

Gibbon, D., Moore, R., Winski, R. (Eds) (1998). *Spoken Language System and Corpus Design*. Berlin: Mouton de Gruyter.

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

Jelínek, T. (2008). Nové značkování v Českém národním korpusu. *Naše řeč*, 91(1), pp. 13–20.

Kopřivová, M., Waclawičová, M. (2005). Construction of Spoken Corpus Based on the Material from the Language Area of Bohemia. In R. Garabík (Ed.), *Computer Treatment of Slavic and East European Languages*. Bratislava: Veda, pp. 137–140.

Křen, M., Waclawičová, M. (2011). Database Framework for a Distributed Spoken Data Collection Project. In S. Goźdź-Roszkowski (Ed.), *Explorations across Languages and Corpora*. Frankfurt am Main: Peter Lang, pp. 83–93.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Miller, J., Weinert, R. (1998). *Spontaneous Spoken Language. Syntax and Discourse*. Oxford: Clarendon Press.

Sgall, P., Hronek, J., Stich, A., Horecký, J. (1992). *Variation in Language. Code Switching in Czech as a Challenge for Sociolinguistics*. Amsterdam: John Benjamins.

Waclawičová, M. (2007). Spoken Corpus ORAL2006, Information It Provides and General Characteristics of Spoken Text. In J. Levická & R. Garabík (Eds), *Computer Treatment of Slavic and East European Languages*. Bratislava: Tribun, pp. 283–289.

Waclawičová, M. (2009). Regionální mluva v korpusu mluvené češtiny ORAL2006 se zaměřením na situaci v českém pohraničí. *Naše řeč*, 92(2), pp. 72–86.

Waclawičová, M., Křen, M. (2008). ORAL2008: New Balanced Corpus of Spoken Czech. In *Труды международной конференции "Корпусная лингвистика – 2008"*. Издательство СПбГУ, pp. 105–112.

Waclawičová, M., Křen, M., Válková, L. (2009). Balanced Corpus of Informal Spoken Czech: Compilation, Design and Findings. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association INTERSPEECH 2009*. Brighton: ISCA, pp. 1819–1822.