Olomouc Corpus of Spoken Czech: Characterization and Main Features of the Project

Petr Pořízka (Olomouc)

Abstract

This study presents the results of the author's research project called Olomouc Corpus of Spoken Czech (OCSC). The paper is focused on the state and partial phases of constructing the corpora, its methodology and annotation. Within the OCSC we use so called dual system of transcription, which means (1) an orthographic one with the purpose of linguistic (morphological) analysis and tagging and (2) a phonetic version of transcript which consists of three layers of the text: first the real transcription and further various types of the metatexts as a second and third layer, including communication aspects of the texts. The criteria of selection of speakers are also listed here and the highly important statistical analysis of the sociolinguistic categories (gender, age, type of education, types of recordings) is presented as well. This analysis can serve as a base for a partial correction of possible non-balance among those sociolinguistic parameters. The annotation rules and principles are mentioned at the end of this study.

1 Introduction

The research project of *Olomouc Corpus of Spoken Czech* (OCSC) is systematically built by the author of this paper from 2002 to date at the Department of Czech Studies at Palacký University in Olomouc (Czech Republic), Faculty of Arts. OCSC, which is pursued as a general corpus typologically, is currently the biggest corpus of spoken Czech (circa 1,5 million words – see table 1). All previous spoken corpora – Prague (PSC) and Brno (BSC) Spoken Corpus as well as ORAL2006 and ORAL2008 – have been constructed on the same methodological base with the modifications at the Institute of Czech National Corpus.[1] (There is one more corpus focused on spoken form of Czech language – specialised corpus DIALOG that is focused on analysis of dialogues in media).[2] The OCSC project started in 2002 (or 2003 respectively) and was firstly based on general methodology of spoken corpora of Czech National Corpus (CNC). We needed to modify and change some methodological aspects because the conception of spoken part of Czech National Corpus is based prevailingly on orthography that doesn't reflect some substantial aspects of spoken language in general.

We've decided to make the changes and modifications based on specificities of the spoken language so radically that we created Czech spoken corpus based on the new conception: we pay close attention to transcription, annotation, format of transcripts, and an appropriate software for processing, managing corpus and querying the data from the corpus (data retrieval).

[1] For further information see http://www.korpus.cz (English version of the web site is available).

[2] Information about this project can be found at http://ujc.dialogy.cz (an English version of the web site is available as well).

spoken corpus	corpus construction period	number of recordings	number of speakers	corpus in size (words)
PSC	1988–1996	304	504	**674 992**
BSC	1994–1999	250	294	**596 009**
ORAL2006	2002–2006	221	754	**1 000 798**
ORAL2008	2002–2007	297	995	**1 000 097**
OCSC	**2002–to date**	**289 (578)**[3]	**658**[4]	**circa 1 500 000**[5]

Chart 1: Corpora of Spoken Czech: CNC and OCSC[6]

2 Fundamentals of Data Collection and Characteristics of Speakers

Data (recording) collection of OCSC is based in accordance with corpora of CNC on the combination of four sociolinguistic variables characterizing speakers: (1) sex (male–female); (2) age (older–younger; with the lowest limit being c. 20 years of age and the limit that is set on 35 years of age); (3) education (lower–higher) and (4) language data that are gained from driven and non-driven way of recording process. For (4): It means (4a) formal recording as a monologue, by course of predefined and thematically wide questionnaire, and (4b) informal recording, which means a non-driven dialogue among speakers (knowing each other well). The informal recording or dialogue is not thematically specialised; the length of recording is set at about 20 minutes (roughly 2 300 words). The optimal number of speakers in dialogue is two or three participants in order to avoid it becoming intelligible due to simultaneous speech. One of the participants was usually also a respondent in the formal recordings, which enables us to observe the differences between the Czech language used in unofficial and semi-official situations.

From the beginning of creating Czech spoken corpora there is a principle that participants recorded are either native speakers of a given area, in this case in Olomouc, or have lived in this area for at least 20 years. In OCSC the rules are not so strict: it is not necessary to be a native speaker, nor to live in Olomouc for at least 20 years. It is essential that a speaker lives or has been living in Olomouc, or he/she has an employment here and comes to Olomouc daily (daily contact with the language variety in Olomouc). We exclude the language of adolescent youth of a given area. Characteristics of participants in OCSC are completed by information about (i) profession, (ii) factual age, (iii) time spent in Olomouc (if the participant does not come from Olomouc), and (iv) region of childhood residence. For purposes of corpus data retrieval we use more detailed age categorization (at least by decades), as the existing segmentation of age category into two values with its age limit (see above) is insufficient in consideration of a sociolinguistic analysis by means of a search engine. The category of edu-

[3] We count a set of formal and informal (FOR+INFOR) recordings as one item, otherwise the number would be doubled, i.e. 578! (Within PSC and BSC are FOR and INFOR recordings counted as two items/files separately.) See also here a section Type of Recording. The definite number of recordings, or transcripts respectively, that will be released officially, can slightly change depending on final and careful selection. The following aspects will be considered: technical quality of recording, authenticity of communicative situation and speakers' language locution, if participant corresponds to the criteria of selection etc. First pre-selection has already been made, for we've had 300 (resp. 660) recordings within OSCS in September 2007.

[4] The definite number of speakers in officially released OSCS can be slightly modified by a final selection of speech recordings, see note 4.

[5] A corpus of half-million in size needs circa 50 hours of speech recordings. One set (FOR+INFOR) of recordings in OCSC lasts 30 minutes on average (circa 21 minutes for INFOR and 8,3 minutes for FOR recording), which means at about 145 hours in total, i.e. 1,5 million words by estimation.

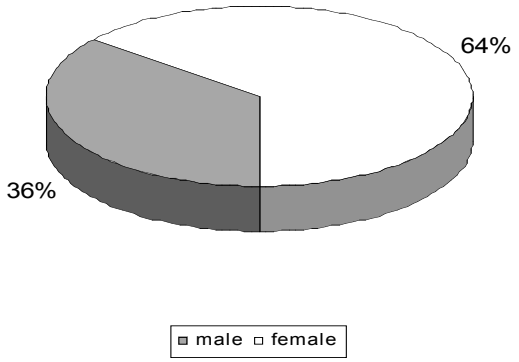
[6] Stadium of development: November 2008.

cation is divided from the original two-values subdivision (BASIS vs. ALTUS) into a trichotomy: (1) primary (BASIS – B), (2) secondary (MEDIUS – M), and (3) university (ALTUS – A) education (see below).

3 Statistic Analysis of Sociolinguistic Variables in OCSC

The statistical analysis of sociolinguistic variables (gender, age, education, type of recording) within all corpus data has been provided, first of all to find out if the corpus is balanced and eventually to appoint disproportions among objects in view. An achievement of balanced input data in spoken corpora is always a very problematic task.[7] As a very important and essential fact we are considering the possibility of bringing into effect an additional "correction" or partial revision of particular (non)-balanced sociolinguistic variables on the base of statistical data. Therefore we'd provide subsequent collection of recordings aimed at the most noticeable disproportion of particular sociolinguistic parameter (gender, age, education).

3.1 Gender of Speakers

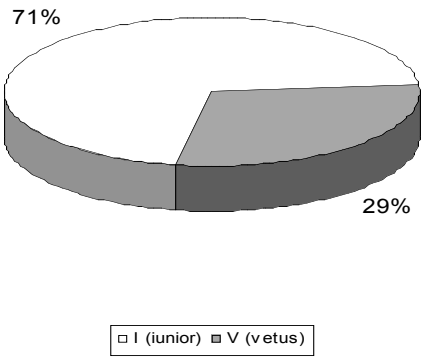


gender	number of speakers
male	236
female	422
total	**658**

Chart 2

The results reveal a marked domination of females. Gender category can be amended with relative ease by subsequent collection of recordings in which men would prevail.

3.2 Age of Speakers



I (iunior) = under 35 years
V (vetus) = above 35 years
Having a balanced corpus in accordance with age of participants it is important to provide statistical analysis on the base of particular age of speakers (or at least by decades). We currently prepare the data for this analysis.

The numbers show that the age category in our corpus is unbalanced. Hence we have also explored a mutual connection between age and gender category, i.e. we've explored the participation of categories IUNIOR–VETUS separately for men and women to get more precise

[7] Statistical analyses are presented by a graphical chart type (percentage ratio), and by a numerical chart.

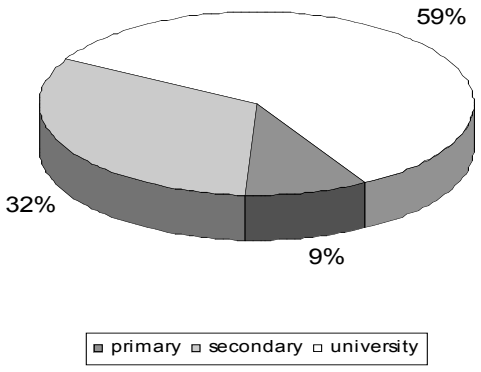
information and to find out the biggest disproportion. The results are listed in Chart 3 (numbers mean a percentage ratio).

	male (%)	female (%)
IUNIOR	28	44
VETUS	8	20
total	**36**	**64**

Chart 3

3.3 Education of Speakers

Previous Czech spoken corpora use the subdivision into two values: BASIS vs. ALTUS – as mentioned above, whereas the term *basis* covers both primary and secondary education. Our three-value subdivision covers following subcategories: BASIS = primary and apprentice education, MEDIUS = secondary one, and ALTUS = commenced, unfinished and finished university education.



education	number of speakers
BASIS	61
MEDIUS	212
ALTUS	385
total	**658**

Chart 4

education	number of speakers	percentage ratio (%)
BASIS	273	41
ALTUS	385	59
total	**658**	**100**

Chart 5

3.4 Type of Recording

The question is, if formal and informal recordings should be considered as two separate types of discourse (note: that they have no common denominator differing form each other), or if it is more suitable and adequate to think of these types of recordings as the only one discourse. Spoken corpora of CNC represent the first concept (see Chart 7). Based on the fact that formal and informal types of recordings are allied together by their methodological matter we find it more suitable to consider FOR and INFOR recordings as just one set connected by one speaker, who takes part in both types of recordings (see Chart 6). It's apparent from Charts 6 and 7 bellow that these two different approaches markedly affect the results of the analysis, especially in case of one, two, or three speakers respectively within one recording of given corpus.

FOR+INFOR as one file/item	number of speakers	percentage ratio (%)
1 speaker	***2***	***0,6***
2 speakers	***211***	***73***
2 speakers	***70***	***24,2***
4 speakers	4	1,4
5 speakers	1	0,4
7 speakers	1	0,4

Chart 6

FOR and INFOR as two files/items	number of speakers	percentage ratio (%)
1 speaker	***291***	***50,3***
2 speakers	***211***	***36,5***
3 speakers	***70***	***12,1***
4 speakers	4	0,7
5 speakers	1	0,2
7 speakers	1	0,2

Chart 7

4 Fundamentals of Mark-Up and Annotation

When creating a conception of mark-up and annotation of OCSC we set several principles considered by us as relevant for spoken corpora in general:

- to choose or develop an adequate method of notation for visualising of audio recordings of spoken Czech, and to decide on punctuality of transcription
- to solve so-called simultaneity of speech-turns of particular speakers, and to put into transcript all relevant communication aspects of dialogues (commentaries and metatext information)
- to differentiate distinctly particular levels of transcription for the purpose of distinguishing the factual text and metatext
- to write down all contextually relevant aspects of dialogues the way that doesn't disrupt a transcript of speakers' utterances
- to create preferably well-arranged and unambiguously structured system of annotation, generally true
- carefully consider a choice and a manner of extra-linguistic and intra-linguistic annotation system

5 Transcription of Recordings: Multilayer Transcript and SVIFT format

Contrary to corpora of CNC we don't use quasi-orthographic type of notation, because the transcription rules of this notation are based markedly and prevailingly on orthography and don't reflect majority of substantial aspects of spoken language. Authors of the quasi-orthographic notation were motivated by requirements of a subsequent morphological analysis and tagging, but to date the corpora of CNC are still not morphologically annotated.

Based on the fact that OCSC is a spoken corpus we've tried to develop such system of notation and transcription that could lead towards an adequate visualisation of a phonetic realization of a speech continuum, and could enable (semi)automatic linguistic annotation by means of some software as well. This is an ambivalent situation: on one hand there is a need to have preferably the most accurate written record of audio-recording, on the other hand the written record should enable technical processing of text. We solved this situation by using a dual system of transcription, which means (1) an orthographic one with the purpose of linguistic (morphological) analysis and tagging, and (2) a phonetic version of transcript that reflects all important aspects of spoken variety of a given language, i.e. Czech, as well as communication aspects of the dialogues (see bellow).

Common text editors are used to create transcripts that are saved into plain text format (*.txt*). For such purposes we've developed a special transcription format called SVIFT (Structural

Vertical and Interlinear Format of Transcription). This SVIFT format enables to execute the automatic conversion into XML format (an international standard for structured data) in the next stage of implementation of corpus data, namely by means of a script written in some scripting language (Perl, Python, etc.).

The defined structural symbols of SVIFT format mark a type of (meta)text: whether it is a factual text of transcript, or a new section, commentary, time reading etc. They always precede an each separate line of transcript (separated by *Enter*), i.e. these signs stand at the beginning of each speech-turn and of metatext lines. The structural symbols are combined with transcription (meta)symbols that are instrumental to mark the simultaneity of speech-turns, commentary sections, an indication of incomplete words, an unintelligible part of recording and other subjects.

Phonetic transcription is therefore multilayer and in comparison to orthographic one it is much more detailed having three layers of text: the real transcription as the first and a basic layer and other various types of metatexts as a second and third layer. The second layer is aimed to structure the text (topic sections followed by time reading) and the third layer serves to capture all metatext information enclosed within commentary (angle) brackets including communication aspects of texts, commentaries, non-verbal and paraverbal events. The particular layers are marked by the dollar sign ($) – the first layer with the phonetic record, the number sign (#) – the second layer with the orthographic record, and the paragraph sign (§) – the third layer with the orthographic record. (Meaning of these signs see also in section *(Meta)Symbols of Annotation – Overview*.)

Important metasymbols are marked by square brackets. They are used to enclose the parts of speech-turns that are realized simultaneously by two (or more) speakers at the same time and signalize the start and the end of overlapping. There is relatively a common situation in dialogue when one speaker enters into the speech of another speaker several times during the only one speech-turn. These square brackets are therefore matching with numerical index (see an example below):

Example:
A: [1 not that we had made arrangements]1 no but / [2 no we are co-debtors]2 but we [3 made arrangements cause romca paid]3 much more than me // i think it's split equally half-half because we have a bond / half-half hey / even though roman's repayments are higher or he pays for both of us
B: [1 it is / there is only one debtor ↑ /]1
B: [2 / it's better /]2
B: [3 you have a share in it based on amount invested ↑ /]3

The use of indexing square brackets serves as an instrument signalizing and identifying the mutual parts of different speakers' speech-turns involved. It's a relevant element of transcripts and has to be marked consistently.

The list of all symbols complemented by marks for prosodic level of utterances (and a short sample transcript) are itemized below.

6 (Meta)Symbols of Annotation – Overview[8]

6.1 MEANING OF STRUCTURAL METASYMBOLS

$	a factual text of transcript (preceding a sign of speaker), i.e. "$A:"
#	commentary, time reading, other metatext information, ex. # < coughing >
§	section (topic) of dialogue/transcript; mark for the change of topic during a dialogue, ex. § Buying of a new car

6.2 METASYMBOLS OF TRANSCRIPTION

[ ]	simultaneity of speech-turns (numbered successively); square brackets are used to enclose the parts of speech-turns that are realized simultaneously by two (or more) speakers at the same time; they match with numerical index successively
( )	words or sequence of words, in case the transcribers are not sure of the real wording (form) (i.e. an influence of worse quality of recording, or a certain deformation of words by the speaker); round brackets are used to enclose any word or sequence of words where the transcriber is uncertain about the correct transcription
< >	commentary (angle) brackets for enclosing of metatext lines (if there are more than one commentary to one line/speech-turn, they are numbered successively)
+	an indication of incomplete word, missing speech sound, or cluster of speech sounds
---	an unintelligible part of recording (difficult to understand for reasons of limited audibility)
:	an inappropriate lengthening of vowels
_	continuous, geminated pronunciation of speech sounds (i.e. *musím _ míd _ dúvot* :: *I have to have a reason*)
@	hesitation or sounds of response (if the hesitation is longer, the sequence of two or three "at" symbols are used, i.e. @@@)
:-)	laugh(ing); graduated by relative "intensity" or length, i.e. :-) :-)) :-))); it is possible to use so-called smileys ☺

6.3 PROSODIC LEVEL SYMBOLS

/ // ///	pause (graduated by relative length of pause, its duration)
´	an emphasis, marked stress/accent (i.e. *stěžoval si na ´tebe* :: *he´s complained about ´you*)
↑ ↓ →	arrows signalize the types of intonation

[8] In several aspects we've been inspired by so-called *Göteborg Transcription Standard*. Overall it's an extensive system of annotation and that's just partly too (and unnecessarily) complicated: some symbols are even doubled. For common practises and majority of purposes this large and extensive annotation system „burdens" the particular transcript too much.

6.4 Sample Transcript

§ A man and his dog
# < time: 00:13:50 >
$B: < do something ↓ hey ↑ >
< a dog >
$A: good ↑ < did you eat it yet ↑ >
< asking his dog >
$B: did you eat it yet ↑ but it doesn't matter ↓ / here it comes there is some wine ↓
$A: <1 was is tasty ↑ >1 / <2 look ↑ now he will be begging ↓ / look ↓ >2
<1 asking his dog >1
<2 pointing at the dog >2
$B: some advert ↑
$A: yeah → it was in the newspapers ↓
… … …

§ Grandmother's party – "tasting"
# < time: 00:17:25 >
$A: < wait ↓ let's taste it ↓ ok ↑ > / it's for our guys anyway ↓
< spiced nuts >
$B: here you are ↓ /// so i don't know ↓ /// it can't be taken out nicely ↓
$C: < some orange flavour ↓ >
< they're eating chocolates of various flavours >
$A: grandma → this is again the most embarrassing what you have ↓ isn't it ↑
$B: oh my god → my colleague yesterday → / he indulges in eating those ninety-nine per cent chocolates →
$C: i wouldn't eat it ↓
$B: but you know what ↓ / that chocolate ↓
$A: i did taste it ↓ / you don't tell a difference ↓
$B: is it worth ↑ it if you can't tell the difference ↑

We are currently developing the tool for an automatic conversion of transcripts into the structured XML format. The fact that all transcripts saved in plain text format (.txt) are structured on the base of SVIFT system enable this conversion.

Simultaneously we are busy with a development of web pages of the whole project, where interested persons will find all information about Olomouc Corpus of Spoken Czech at *www.corpus.upol.cz*.

References

Czech National Corpus – Prague Spoken Corpus, 2001. Institute of Czech National Corpus, Charles University in Prague, Faculty of Arts. Available at http://ucnk.ff.cuni.cz.

Czech National Corpus – Brno Spoken Corpus, 2001. Institute of Czech National Corpus, Charles University in Prague, Faculty of Arts. Available at http://ucnk.ff.cuni.cz.

Czech National Corpus – ORAL2006, 2006. Institute of Czech National Corpus, Charles University in Prague, Faculty of Arts. Available at http://ucnk.ff.cuni.cz.

Czech National Corpus – ORAL2008, 2008. Institute of Czech National Corpus, Charles University in Prague, Faculty of Arts. Available at http://ucnk.ff.cuni.cz.

Nivre Joakim et al. (2005): *Göteborg transcription standard. Semantics and Spoken Language, Version 6.3 – DRAFT*. Göteborg University. Available at http://www.ling.gu.se/projekt/tal/doc/transcription_standard.html.

Olomouc Corpus of Spoken Czech (meantime unpublished; Dept. of Czech Studies, Palacký Univesity, Faculty of Arts)

Pořízka Petr (2004a): "Představení výzkumného projektu Olomoucký korpus mluvené češtiny". In: *AUPO Moravica I*, Olomouc, VUP: 125–131.

Pořízka Petr (2004b): "K možnostem vyhledávání dat a struktuře atributů korpusových manažerů v mluvených korpusech ČNK". In: *AUPO Philologica 84, Bohemica IX*, Olomouc, VUP: 81–92.

Pořízka Petr (2005): "Přepis(y) textů v korpusech mluvené češtiny". In Pořízka, Petr/Polách Vladimír (eds.): *Jazyky v kontaktu/jazyky v konfliktu a evropský jazykový prostor*, Olomouc, VUP: 235–240.

Pořízka Petr (2008): "Anotace orálních korpusů. Olomoucký mluvený korpus jako model". In: Kopřivová, Marie/Waclawičová, Martina (eds.): *Čeština v mluveném korpusu*. Praha, NLN: 177–189.

Pořízka Petr (2008): "Olomoucký mluvený korpus – stav, metodologie, charakteristika". In: Štícha, František/Fried, Mirjam (eds.): *Grammar and Corpora / Gramatika a korpus 2007*. Praha, Academia: 191–198.