Mapping Diatopic and Diachronic Variation in Spoken Czech: the ORTOFON and DIALEKT Corpora

Introduction

ORTOFON and DIALEKT are two corpora of spoken Czech currently in preparation at the Institute of the Czech National Corpus (CNC). They will make available, respectively, the most recent and the oldest systematic recordings of the language focusing on the broadest possible coverage of the territory of the Czech Republic. Even though the methodologies of data collection and annotation differ to a certain extent between the two corpora (see Tab. 1 for overview), we hope to enable a diachronic analysis of the changes which have occurred within spoken Czech in the last 50 years, as well as an assessment of synchronic diatopic (region-based) variation.

	ORTOFON	DIALEKT
will be completed by	end of 2016	
target size (tokens)	1,000,000	200,000
target size (hours of audio)	110 (est.)	22 (est.)
current size (raw data)	~ 900,000 tokens / 97 hrs	~ 30,000 tokens / 2.5 hrs
transcription tiers (1 of each per speaker; see Fig. 1)	orthographic, phonetic and metalinguistic	dialectological, orthographic and metalinguistic
time-frame of data collection	2012–ongoing	1960s–1980s
type of material	non-scripted informal interactions	mostly monologues, narratives
recording methodology	covert recording	dialectological interview
social coverage	broadest possible	older rural speakers
regional coverage	broadest possible	traditional dialect areas
age coverage	> 18 y.o.	mostly > 60 y.o.
format	audio aligned with multi-tier transcript	

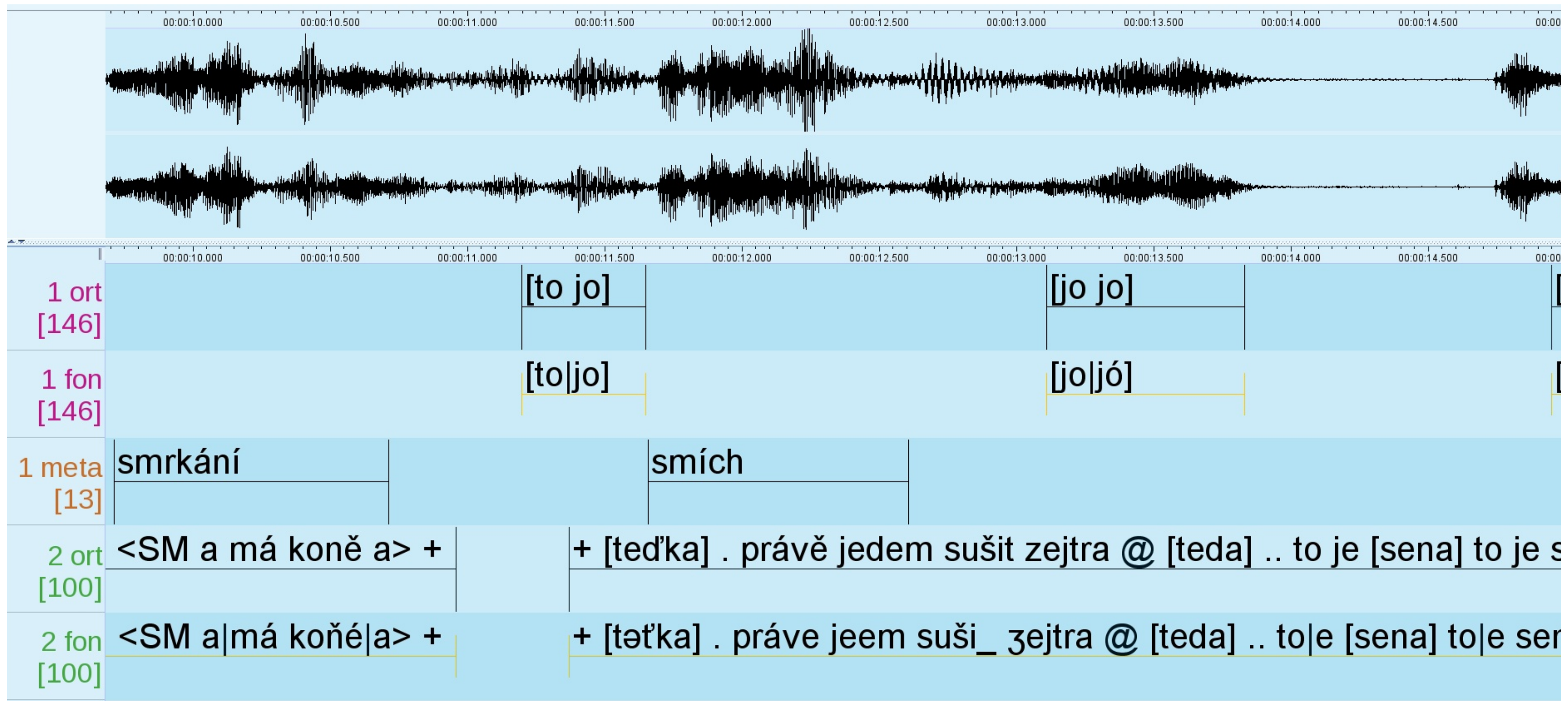
Tab. 1. The ORTOFON and DIALEKT corpora at a glance. NB: A token is defined as a position in the corpus containing alphabetic characters.

Annotation Scheme

The recordings for both corpora are transcribed using a multi-tier annotation setup implemented via the ELAN linguistic transcription software. There are two main types of tier and each speaker in the conversation gets his or her own private instance of both of them, which means that any overlaps may be conveniently transcribed in parallel on the respective independent layers. Speakers' turns are segmented into sub-units of a maximum length of 25 tokens.

In ORTOFON, the first tier carries a transcript which mostly sticks close to Czech orthography, enriched with selected phonetic and lexical regional variations. False starts, pauses and hesitations are also marked, as well as the boundaries of overlapping speech. The second tier uses a simplified and adapted form of phonetic transcription, which will make it possible to assess quantitatively various features of spoken Czech: assimilations, vowel reductions, stress groups, cliticization. Alongside the two main tiers (orthographic and phonetic), auxiliary metalinguistic layers also capture concomitant acoustic events such as non-verbal or ambient sounds. For an example, see ELAN screenshot in Fig. 1.

In DIALEKT, the first tier is a dialectological layer and the second an orthographic one (corresponding to the one in ORTOFON). The dialectological layer is transcribed according to well-established rules for transcription in fieldwork on varieties of Czech: the set of symbols used is a superset of the Czech alphabet, which makes it possible to capture speech sounds characteristic to non-standard varieties, but the word boundaries are kept intact (as in standard written language) and conventional punctuation is used.



1 ort [146] [to jo] [jo jo]

1 fon [146] [to|jo] [jo|jó]

1 meta [13] smrkání smích

2 ort [100] <SM a má koně a> + + [teďka] . právě jedem sušit zejtra @ [teda] .. to je [sena] to je s

2 fon [100] <SM a|má koňé|a> + + [təťka] . práve jeem suši_ ʒejtra @ [teda] .. to|e [sena] to|e ser

Fig. 1. Excerpt from a transcript for the ORTOFON corpus in the ELAN transcription program, showing the recording waveform at the top with various time-aligned transcription layers for two speakers. Please ask if interested in transcription details.

The ORTOFON Corpus

Our current data collection project for the ORTOFON corpus builds on know-how established during a series of projects starting in 2002, during which several corpora of informal spoken language (the ORAL series) have been designed and built at the Institute of the Czech National Corpus. For each recording, we track a variety of metadata spanning the two broad categories of situation and speaker characteristics. In addition to pre-defined categories, some entries are augmented with a free-form specification field, so that the most precise information possible be recoverable.

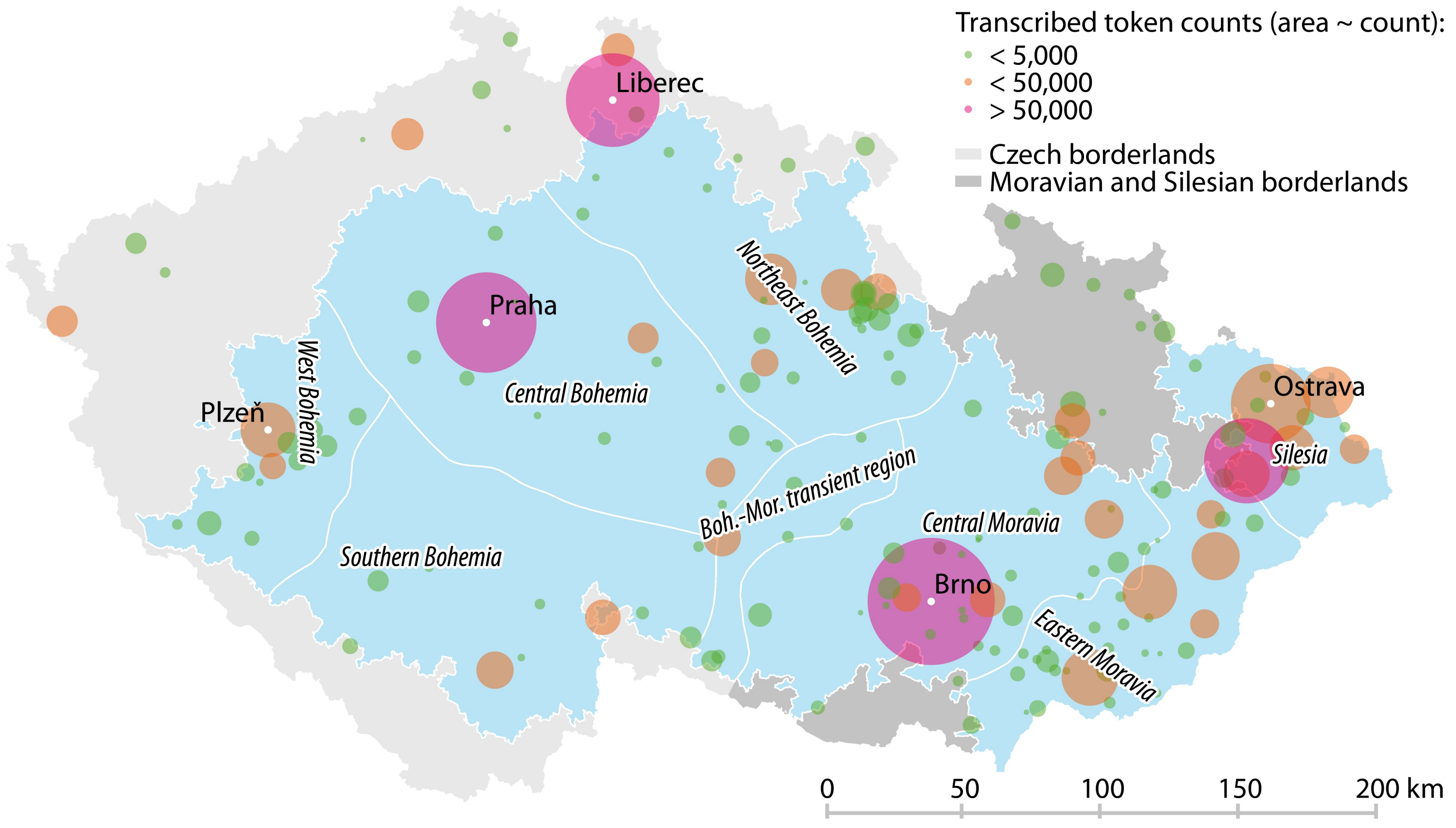


Fig. 2. Regional distribution of amount of tokens transcribed so far for the ORTOFON corpus on the orthographic layer. The 8 traditional dialect regions of Czech, along with two mixed dialect borderland areas, are also indicated.

Situation type

Major conversation topics are summarized (using free-form keywords) and relationships between the speakers are specified (one of partners, family, friends, acquaintances or strangers), as well as the total number of generations they represent (e.g. a child, her mother and her grandmother = 3 generations).

Additionally, several basic situation types to pick from are provided, including:

1. at home (e.g. during a collective activity)
2. public transportation
3. informal chat at work/school
4. visit or celebration
5. restaurant or pub
6. tabletop, RPG or similar game
7. phone or VoIP conversation
8. garden or cottage conversation

Main speaker characteristics

- sex and age
- education level and field
- current and longest occupation
- common speech defects
- childhood, longest and current region and place of residence, and size of the corresponding dwelling

The region of residence category is structured according to traditional dialect regions as outlined in Fig. 2.

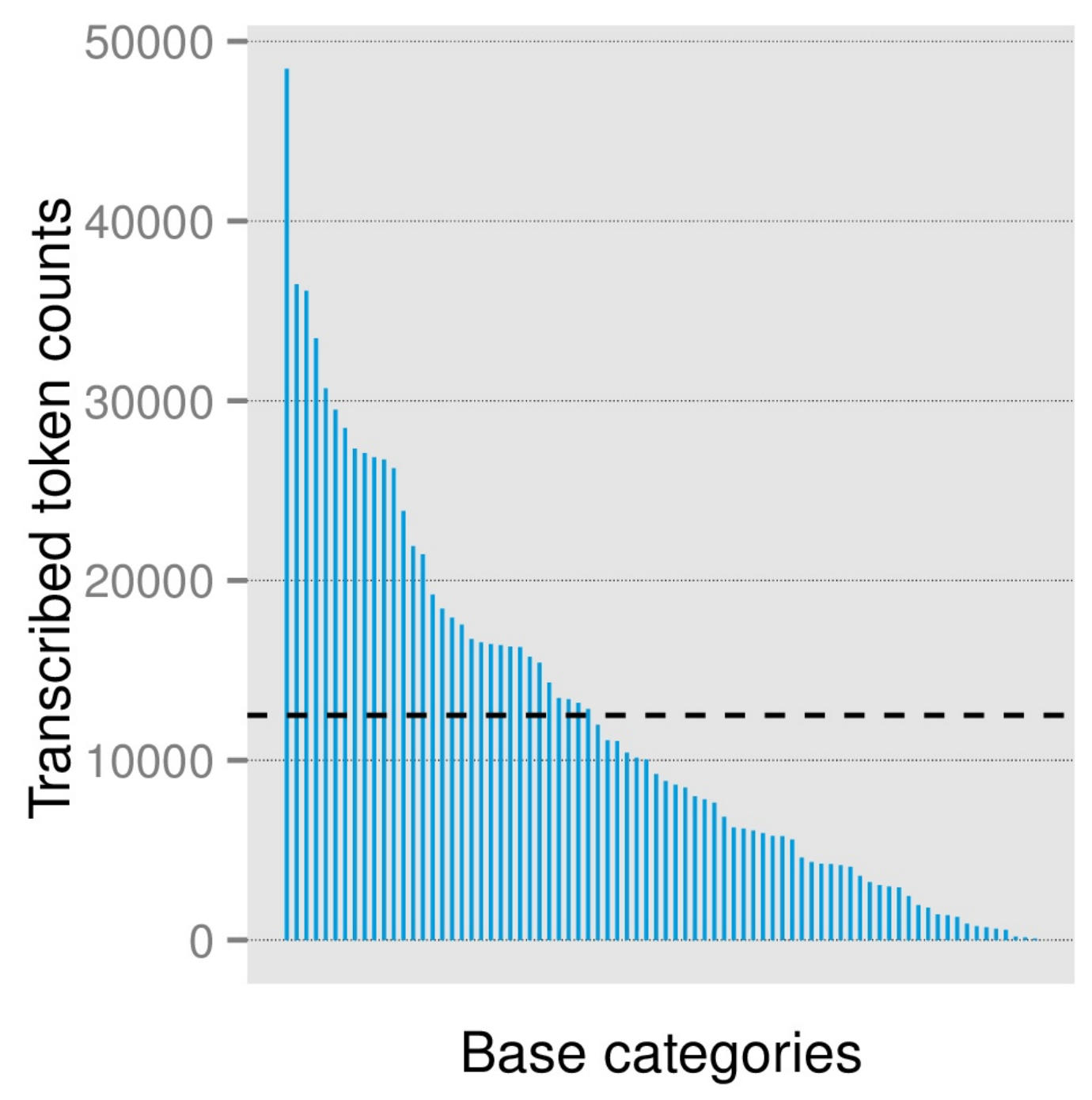


Fig. 3. Counts of tokens transcribed so far for the ORTOFON corpus on the orthographic layer by base sociological categories (see text). The horizontal dashed line indicates the ideal 12,500 token count per category in the final corpus.

Following previous practice (our ORAL series of spoken corpora), we have defined a smaller subset of the above-mentioned categories, collapsing some of them in the process into larger bins, based on which material will be selected for inclusion into the final 1M-token balanced corpus:

- sex and age bin (under 35 y.o. × over 35 y.o.)
- highest attained education level bin (tertiary × non-tertiary)
- childhood region of residence (10 categories; see Fig. 2)

This design results in 2 × 2 × 2 × 10 = 80 base categories, i.e. a target count of 1,000,000/80 = 12,500 tokens per category. In the first stage of data collection, collaborators were allowed to contribute recordings freely (see over-represented categories in Fig. 3); we are currently explicitly targeting under-represented groups.

The DIALEKT Corpus

The DIALEKT corpus is based on recordings of Czech dialects made mainly from the 1960s to the 1980s. The material in the corpus is therefore highly interesting from a diachronic point of view, because it is a repository of archaic dialectal features from regional varieties of Czech, which have mostly become extinct in prevalent contemporary usage. The collected sound material mainly features monological accounts in informal settings (at home), with topics revolving around agriculture, crafts, local customs and traditions, and everyday country life. Features from the individual dialect areas and from all levels of linguistic analysis (phonetics, phonology, morphology, syntax and lexicon) have been captured in these accounts and the DIALEKT corpus will allow to search for them. Once completed, the DIALEKT corpus should be representative primarily in two respects: it should reflect 1) all traditional dialect areas of the Czech Republic, as well as 2) all dialectologically relevant features from the individual areas.

The interface to the DIALEKT corpus will thus include interactive dialect feature maps covering the individual regional varieties and samples of recordings and transcriptions from selected locales. It is planned that data from the ORTOFON corpus will be made accessible through this map in the future as well, using their location metadata. The goal is to enable a comparison of the spread of various dialectal features (e.g. [v]-prothesis, [é]-raising, various kinds of assimilations) in both space and time.

Acknowledgements

This presentation and the associated data collection project is a product of the implementation of the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

Authors: Marie Kopřivová, Hana Goláňová, Petra Klimešová, David Lukeš
Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague, Czech Republic
Contact: {marie.koprivova,hana.golanova,petra.klimesova,david.lukes}@ff.cuni.cz

CZECH NATIONAL CORPUS