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

Marie Kopřivová, Hana Goláňová, Petra Klimešová, David Lukeš
Institute of the Czech National Corpus, Faculty of Arts, Charles University
Panská 7, 110 00 Praha 1, Czech Republic
{marie.koprivova,hana.golanova,petra.klimesova,david.lukes}@ff.cuni.cz

Abstract

ORTOFON and DIALEKT are two corpora of spoken Czech (recordings + transcripts) which are currently being built at the Institute of the Czech National Corpus. The first one (ORTOFON) continues the tradition of the CNC's ORAL series of spoken corpora by focusing on collecting recordings of unscripted informal spoken interactions ("prototypically spoken texts"), but also provides new features, most notably an annotation scheme with multiple tiers per speaker, including orthographic and phonetic transcripts and allowing for a more precise treatment of overlapping speech. Rich speaker- and situation-related metadata are also collected for possible use as factors in sociolinguistic analyses. One of the stated goals is to make the data in the corpus balanced with respect to a subset of these. The second project, DIALEKT, consists in annotating (in a way partially compatible with the ORTOFON corpus) and providing electronic access to historical (1960s–80s) dialect recordings, mainly of a monological nature, from all over the Czech Republic. The goal is to integrate both corpora into one map-based browsing interface, allowing an intuitive and informative spatial visualization of query results or dialect feature maps, confrontation with isoglosses previously established through the effort of dialectologists etc.

Keywords: spoken corpus, sociolinguistics, Czech

1. Introduction

This paper introduces ORTOFON and DIALEKT, two corpora of spoken Czech which are currently in preparation at the Institute of the Czech National Corpus (CNC) and which will make available, respectively, the oldest and the most recent 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, we hope to enable a comparative analysis of the changes which have occurred within spoken Czech in the last 50 years, as well as an assessment of synchronic (region-based) variation.

2. Corpora of Spontaneous Spoken Interactions

2.1. General Aspects

Though smaller in general than written language corpora, corpora of spontaneous speech are crucial to linguistic research because they are repositories of language in use in its primary medium, with all the specificities and significant differences from more readily available textual resources that this entails. Linguistic phenomena for which these are an invaluable resource include word-formation strategies in spontaneous utterances, lexical tendencies and trends in spoken language (e.g. formation and use of interjections or particles), phonetic aspects of language (connected speech processes, deformation of frequent and/or filler words), and last but not least, diachronic and diatopic variation, as informal spoken language is much less resistant to change than its formal counterpart, not to mention the standardized written variant. This is especially salient with respect to the Czech linguistic situation, which has been argued to be one of diglossia (Hammer, 1985; Čermák, 1993), with important differences between the language expected in formal communication and the varieties (dialects and interdialects) used in everyday interactions.

From the point of view of conversation analysis, balanced interactions of two speakers or more are the most interesting ones, because they offer the most evidence of topic-establishing, topic-switching and turn-taking strategies, as well as patterns which have traditionally been of interest to the field, such as adjacency pairs, pre-sequences and repair sequences (Hutchby and Wooffitt, 2009). Sociology (or more narrowly, sociolinguistics) will study the influence of sociological variables on the unfolding of the interaction and on the linguistic material employed; psychology, the way partners in communication manipulate and/or accommodate to each other; media studies, strategies of persuasion and linguistic means for achieving certain goals (acting out a conviction, swerving a debate in one's favour).

Yet the usefulness of these corpora is not limited to linguistics or more broadly to the social sciences; it is also very real in terms of engineering research and applications. N-gram language models based on these corpora capture the specificities of spoken language and cannot be easily substituted for by models derived from textual data. They have been for instance successfully integrated into the pipeline of semi-spontaneous speech recognition systems, such as one for the automated transcription of lectures in Czech (Mikolov et al., 2008).

2.2. The ORAL Corpora Series

Our current spoken data collection project for the ORTOFON corpus is the newest in a series of spoken language corpora designed and built at the Institute of the Czech National Corpus since 2002. This series comprises the ORAL2006 (Kopřivová and Waclawičová, 2005) and ORAL2008 (Waclawičová and Křen, 2008) corpora, which both contain only transcriptions, and the ORAL2013 corpus (Válková et al., 2012; Benešová et al., 2013), which

was released in late 2013 and which provides access to actual recordings aligned with a one-tier transcript. These corpora focus on capturing spontaneous spoken language in non-scripted interactions, i.e. what Čermák has termed "prototypically spoken texts" (2009, 118). The speakers know each other and appear in their usual roles, with only our associate (the recorder) being aware of the conversation being recorded. The interactions take place in familiar environments (e.g. in private, among friends etc.) and the situations are not experimentally induced. We only record the speech of adult speakers (18+ y.o.).

The ORTOFON corpus will be completed by the end of 2016 and will contain audio recordings aligned with transcripts; its estimated final size is around 1,000,000 tokens.[1] According to prior experience with previous ORAL series corpora, this should roughly correspond to 110 hours of audio. The recordings will span the whole territory of the Czech Republic. Our goal is for them to be sociolinguistically well balanced for the speaker sex, age, education and region of origin variables. The balancing is being done gradually, see section 3.2. Compared to the previous instalments in the ORAL series, a more stringent system has been put in place to resolve speaker identity in cases where chance had it that one and the same speaker was recorded by two or more of our collaborators. This should ensure that the per-speaker limit of 10,000 tokens maximum is not accidentally violated.

3. The ORTOFON Corpus

3.1. Metadata

Our external collaborators who record and transcribe the conversations are asked to provide a variety of metadata along with each recording, spanning the two broad categories of situation and speaker characteristics as outlined by Crowdy (1993), his own terms being "context-governed" and "demographic" perspective. These are fairly detailed and should enable users to filter for specific types of extralinguistic context and, provided that enough material matches their search criteria, to create subcorpora based on them. A coarse-grained subset of this information will be used for the sociolinguistic balancing of the corpus (see section 3.2.).

3.1.1. Situation

Each recording contains information about the situation in which it was made. There is a forced choice of primary situation type from a list of 12 pre-defined categories, which are designed to distinguish, among the different possible settings in which the recording could have taken place, those that are of interest:

1. at home
2. at home during a meal
3. at home during a collective activity
4. public transportation
5. visit
6. informal chat at work/school
7. celebration
8. garden/cottage conversation
9. restaurant/pub
10. on the street/at a public transportation stop
11. tabletop, RPG or similar game
12. phone or VoIP conversation

Situation type #8 in particular may seem overly specific, but it only reflects a typical aspect of the Czech cultural context, in which spending weekends at one's cottage is a favourite pastime. Collaborators may select "other" as situation type if none of the provided ones accommodates their case.

Apart from situation type, collaborators are asked to summarize the major conversation topics (using free-form keywords) and specify the relationships between the speakers (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). Another requirement is to enter the place and corresponding geographical area (based on dialect areas, see section 3.1.2.) of the recording. This is potentially relevant for speakers from the more dialectally diversified regions in the east of the country (Moravia and Silesia, see 1), who in some cases tend to eschew regionally marked variants in their speech when in the more dialectally uniform west (Bohemia). For an interesting study of this phenomenon, see Wilson (2010).

3.1.2. Speaker Characteristics

Only stable speaker characteristics—as opposed to transient ones, see Gibbon et al. (1998, 111)—are being consistently tracked. Apart from sex and age, these include:

- education level (highest achieved) and field
- current and longest occupation
- childhood, longest and current region and place of residence, and size of the corresponding dwelling
- common speech defects

In addition to pre-defined categories under all of these headings, occupation and location entries are augmented with a free-form specification field, so that the most precise information possible be recoverable. For instance, location entries will be used to link recordings with an interactive map allowing selection and playback of material based on geographical criteria (for more detail, see section 4.3.). The region of residence category is not structured according to the current administrative units of the Czech Republic, whose boundaries were in part defined somewhat arbitrarily, but based on traditional dialect regions as defined in Balhar et al. (1992; 1997; 1999; 2002; 2005). These are outlined in fig. 1. Note that the borderland regions (denoted with polka dots and stripes) are problematic from a dialectological point of view (Balhar et al., 2011, 10), as a substantial part of the original population, predominantly German-speaking, was deported as a consequence of post-World War II ethnic cleansing and replaced with Czech-speaking settlers from all over the country (Kastner, 1996).

[1] We define a token as a position in the corpus containing alphabetic characters, i.e. not punctuation for instance.

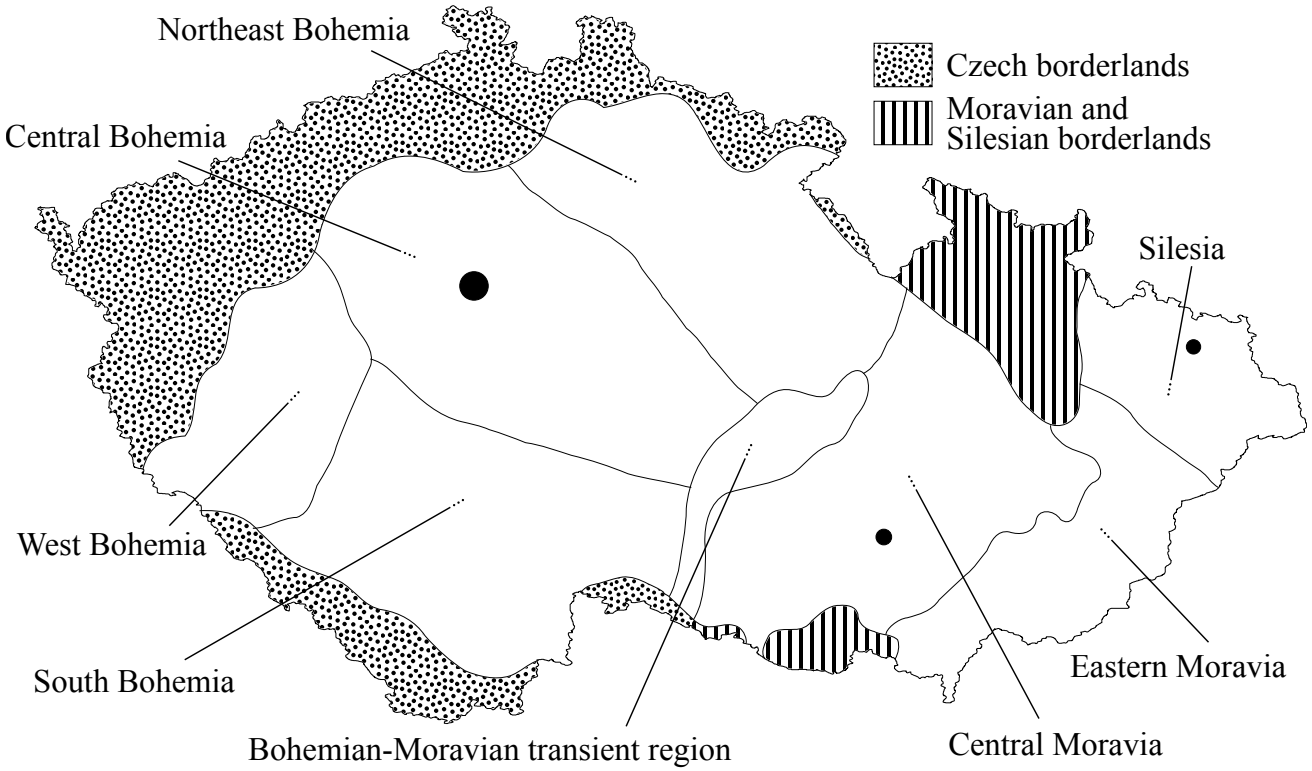

Figure 1: Map of the dialect regions of Czech, following Balhar et al. (1992). The three black circles represent the three largest cities in the Czech Republic, with size roughly corresponding to population; from west to east: Prague, Brno, Ostrava.

3.2. Representativeness and Balancing

Using the metadata sketched out above, users will be able to filter for particular kinds of recordings they are interested in, e.g. conversations involving at least two generations of men who didn't go to university and are friends, provided that the final data set contains such entries. However, it is unrealistic to aim for guaranteeing that all imaginable combinations of these variables will be represented, considering the expected size of the corpus and the laboriousness of data acquisition and transcription. In other words, with respect to all possible combinations of these factor levels, the corpus will be neither **balanced** (i.e. containing roughly equal quantities of data for all these factor level combinations) nor even **representative** (i.e. containing at least *some* data for all of them). Instead, we have defined a subset of these categories, collapsing some of them in the process into larger bins, with respect to which balancing will be attempted:

- sex
- highest attained education level bin (tertiary $\times$ non-tertiary)
- age bin (under 35 y.o. $\times$ over 35 y.o.)
- childhood region of residence (see fig. 1)

This design results in $2 \times 2 \times 2 \times 10 = 80$ base categories, i.e. a target count of $\frac{1{,}000{,}000}{80} = 12,500$ tokens per category, but even so, perfect balancing will inevitably be a daunting task and a hard goal to achieve. It is expected that a redundant amount of data will be collected and consequently selected from,[2] as it is impossible to fully plan the recording sessions ahead of time, but what material will be available ultimately depends on the milieus that our collaborators have access to. As can be seen from fig. 2, which plots current token counts[3] by base category, for some of them, we have been able to secure the cooperation of perhaps even overly zealous collaborators, whilst for others, we are still struggling.

In the first stage of data collection, collaborators were allowed to contribute recordings freely, irrespective of the base categories. Recently, caps have been applied as we are moving into the second stage in which we are explicitly targeting under-represented groups. It is our hope that even should highly exact balancing prove an unfeasible task in the end, the corpus will at the very least be representative (i.e. each category will have a non-zero token count).

In comparison with the target 12,500 tokens per base category, the individual speaker limit of 10,000 tokens may seem rather too benevolent, but let us reiterate that the "raw" data from which the final corpus will be constructed are expected to be redundant, and a higher per-speaker limit gives useful leeway when selecting particular recordings for inclusion. Still, in connection with this, it is unfortunately impossible to say at this point whether it will be manageable to honour across the board the sociolinguistic rule-of-thumb of having at least five speakers per base cate-

[2] The rest of the data will be made available as well, separately.

[3] As transcribed on the orthographic layer of transcription, see section 3.3.

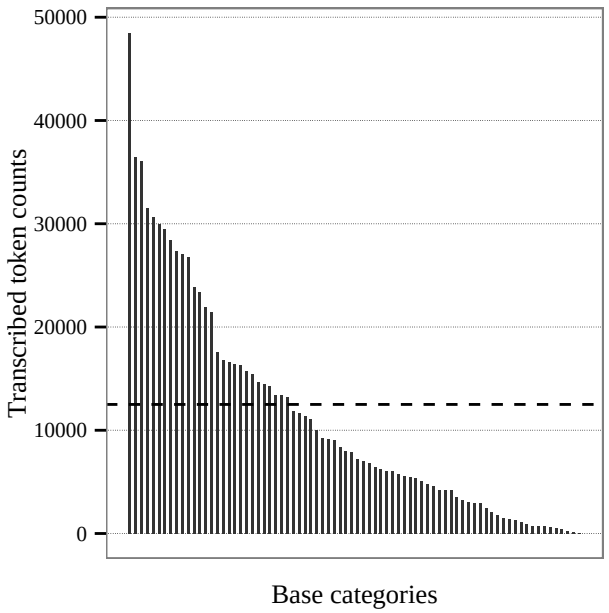

Figure 2: Counts of tokens transcribed so far for the ORTOFON corpus on the orthographic layer (see 3.3. for a characteristic of the transcription scheme), by base sociological categories (see text). Four base categories where we have no speakers as of yet have been omitted. The horizontal dashed line indicates the ideal 12,500 token count per category in the final corpus. In total, the material transcribed so far amounts to 876,859 tokens and roughly 88.5 hours of raw audio material.

gory (Feagin, 2002, 29), which aims to minimize the risk of invalid generalizations with respect to some of the groups delineated by the factors.

3.3. Annotation Scheme

In contrast to the previously mentioned ORAL series corpora, the data in ORTOFON benefit from a multi-tier annotation setup implemented via the ELAN linguistic transcription software[4] (Sloetjes and Wittenburg, 2008). Consequently, a different approach to transcription was employed, compared to the previous instalments in the ORAL series. 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.

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. Conversely, more fine-grained phonetic phenomena like vowel reductions or assimilations are left out. The second tier uses a simplified and adapted form of phonetic transcription, which was designed with the size of the data and accessibility for the corpus user in mind. Basic search and lemmatization will rely on the orthographic layer, but the phonetic layer will be searchable as well. The phonetic transcription will make it possible to assess quantitatively the features of spoken Czech: stress groups, vowel reductions, cliticization.

Alongside the two main tiers (orthographic and phonetic), auxiliary layers also capture concomitant acoustic events such as non-verbal or ambient sounds. Selected paralinguistic aspects of an utterance (e.g. when speech is accompanied by laughter) are recorded directly within the orthographic transcript. Some types of proper names are anonymized both in the recordings and in the transcript. By default only family names are removed, but the recordees can request more information to be excised once they are retrospectively informed of having been recorded.

An example of what the tiered transcript looks like in the ELAN program is provided in fig. 3, including a brief and non-exhaustive explanation of the symbols used (limited only to those actually present in the screenshot). Note in particular that some orthographic words are merged into prosodic words on the phonetic tiers, but the space between them is not simply removed. Instead, it is replaced with the pipe (|) symbol, so as to preserve information about the orthographic boundary location and, by extension, a one-to-one correspondence between the tokens on the two tiers. This will allow search query constraints to target both tiers simultaneously, providing the users with more control over their search results.

3.3.1. Phonetic Transcription: Challenges

Devising an appropriate phonetic transcription system has not been entirely straightforward. Phonetic annotation by hand requires either highly trained transcribers or a vastly simplified transcription apparatus; we opted for a middle road, i.e. a moderately complex transcription system (isomorphic with a subset of the International Phonetic Alphabet) and training our transcribers (some of them already having some background in phonetics) as we go by way of tutorials and extensive feedback. The transcription must thus negotiate between several conflicting requirements:

- to capture as faithfully as possible the real pronunciation (requires detail)
- searchability of the corpus (requires generalization)—it would be cumbersome to make concordances from the phonetic layer if transcriptions were overly individualized and detailed (i.e. if too few tokens ended up with the same or a predictably similar transcription)
- to enable non-phoneticians to pick up the transcription rules quickly as they go (requires simplicity)

The crux of the problem is that simplifying the transcription system means biasing it towards phenomena we are keen on capturing, or simply those we already happen to know about, and sidelining those we care less about or whose existence and/or relevance we ignore as of yet. Yet at the same time, we would like the transcription to remain as faithful and objective as possible, so that the corpus be general

[4] ELAN is being developed at the **Max Planck Institute for Psycholinguistics**, The Language Archive, Nijmegen, The Netherlands; URL: `http://tla.mpi.nl/tools/tla-tools/elan/`

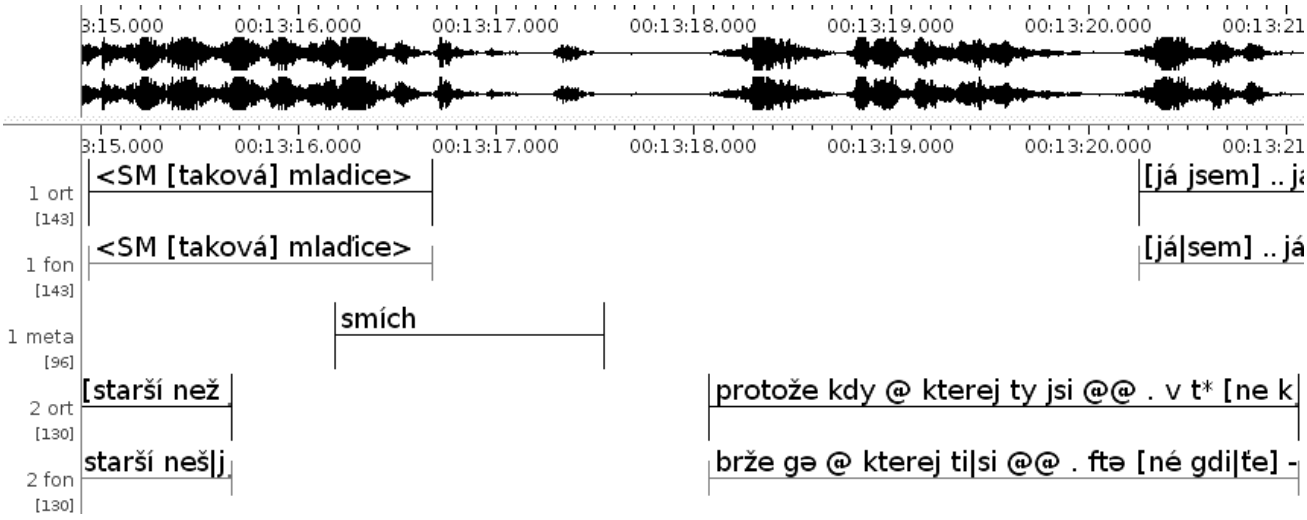

Figure 3: Excerpt from a transcript for the ORTOFON corpus in the ELAN transcription program, showing the recording waveform at the top with time-aligned orthographic, phonetic and metalinguistic tiers for speaker 1 (`1 ort`, `1 fon`, `1 meta`) and orthographic and phonetic tiers for speaker 2 (`2 ort`, `2 fon`); additional tiers not displayed. **Angle brackets** (`<...>`) provide additional information on the way the text is uttered (e.g. the tag `SM` indicates concomitant laughter). **Square brackets** (`[...]`) signal overlapping speech. **Two dots** (`..`) indicate a pause (when longer than 2 s, pauses are annotated separately on the metalinguistic layer), **one dot** points to a salient non-pause prosodic boundary, which is recorded only optionally. **Stars** (`*`) are used for flagging incomplete words (false starts) and the **at-sign** (`@`) for hesitations. The **minus sign** signals an incomplete (interrupted) utterance. Additional symbols used in transcription but not present in this screenshot are not explained; for commentary on the **pipe signs** (`|`), see text at the end of section 3.3.

enough to be useful for studying even phenomena we are not necessarily already aware of.

3.4. Technical Quality

In comparison with our prior spoken corpus projects and primarily because of the phonetic transcription layer, more emphasis has been placed on the quality of the recordings, which is sometimes hard to ensure in non-experimental settings. Though there is an unquestionable improvement in overall sound quality, it remains uneven both for technical (dynamic range of the microphone, issues of placement of the portable recording device) and context-dependent (ambient noise, number of speakers, overlaps) reasons. This makes it impractical to implement batch pre-processing of the recordings using DSP or NLP methods. For instance, a first approximate transcription could in theory be derived using an HMM-based speech recognizer, but in practice, this is made difficult by the nature of the speech material (spontaneous dialogues) and overlaps with other speakers' turns and/or a variety of non-verbal sounds. Admittedly, there exist ways to disentangle the individual sound sources in a mixed signal, such as Independent Components Analysis (see e.g. (Mitianoudis, 2004)); however, ICA requires as input multiple simultaneous recordings (one per sound source to separate), which is impracticable for us. Furthermore, though ICA works well in theory, its success rate on real field-recorded data is debatable.

4. The DIALEKT Corpus

The DIALEKT corpus presents traditional regional dialects as captured throughout the territory of the Czech Republic. It is based on recordings of Czech dialects made mainly from the 1960s to the 1980s, predominantly by the dialectological department of the Institute of the Czech Language at the Czech Academy of Sciences (Balhar et al., 2011). 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. In this respect, the corpus which is being built out of these data is specific and unique.

4.1. Speakers and Material

In order to be able to document Czech dialects in their most distinctive forms (i.e. the earliest available ones, given the levelling trend fostered by audiovisual media and increased mobility), the fieldwork targeted primarily members of the oldest generation, the corresponding archetype in English dialectology being the so-called non-mobile older rural male or *NORM* (Chambers and Trudgill, 1998, 29). The recorded speakers are all local natives from rural areas who have never moved during their lives, which means they belong to the settled stratum of the population bound to a traditional way of life (arts and crafts, farming). They were mostly born at the close of the 19th century or the beginning of the 20th. 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. The recordings have been given names which are indicative of the range of subjects covered, for instance Weaving, The Bewitched Snake, Stealing Is Wrong etc. Dialectal 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.

4.2. Annotation

As with the ORTOFON corpus, an annotation scheme with two main tiers per speaker has been devised for the material, consisting of a dialectological layer (specific to this corpus) and an orthographic one (corresponding to the one in ORTOFON). The fundamental layer is the dialectological one, transcribed according to rules for transcription in fieldwork on varieties of Czech (Dialectological Commission of the Czech Academy of Sciences and Arts, 1951). These rules are well-established practice in Czech dialectology: 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.

The second layer carries the so-called orthographic transcription, which comes close to standard spelling conventions and follows the rules for spoken corpus annotation devised at the Czech National Corpus. This second tier will therefore allow direct comparison with material contained in the other spoken corpora of the CNC.

4.3. Visualization on a Map

When processing linguistic data from a spoken corpus, it is often useful to categorize speakers via a hierarchical system of linguistically relevant geographical units. The macro areas of this system for Czech, as determined and refined over the years by dialectologists on the basis of isoglosses which distinguish among regional varieties, are depicted in fig. 1. A high-accuracy digital version of this map, augmented with the locations of a number dwellings, has been commissioned using the ArcGIS[5] system, which will allow users to locate the recordings with precision and confront them with established dialect boundaries.

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, in the spirit of tools such as `http://www.dialektkarte.de` (König, 2005). It is planned that data from the ORTOFON corpus will be made accessible through this map in the future as well. The goal is to enable a comparison of the spread of various dialectal features (e.g. [v]-prothesis, [eː]-raising, various kinds of asssimilations) in both space and time.

Once completed, the DIALEKT corpus will count roughly 200,000 tokens. It should be representative primarily in two respects: it should reflect 1) all dialect areas of the Czech Republic, as well as 2) all dialectologically relevant features from the individual areas. It is our hope that it will be useful not only as a research tool (for linguists specializing in dialectology as well as other fields), but also as a teaching aid in both secondary and tertiary education. No corpus with this type of historical data on the Czech linguistic situation is currently freely accessible to the public.

5. Conclusion

Together, the ORTOFON and DIALEKT corpora will thus allow users to explore diachronic and diatopic variation in spoken Czech through a convenient integrated interface. Compared to previous spoken corpora built at the Institute of the Czech National Corpus (the ORAL series), they feature a more detailed annotation system separated into several parallel layers accommodating speakers individually. A rich set of both context-dependent and demographic metadata characterizing the recordings provides additional perspectives on the collected material. A problem we are currently intensively working on solving is that of the query and concordance interface to our data, as both our present corpus query engine (Rychlý, 2007) and concordancing graphical web interface,[6] used with the ORAL series corpora as well as with other corpora created at the ICNC, are ill-adapted for retrieving and displaying data from multilayered transcriptions. After a preliminary overview, a tool with capabilities such as those of the ANNIS corpus manager (Zeldes et al., 2009)—mainly in terms of the ability to constrain queries on multiple tiers based on time overlap of segments—appears of necessity.

6. Acknowledgements

This paper 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.

The authors would also like to thank three anonymous reviewers for their helpful comments.

7. References

Balhar, J., Jančák, P., Bachmannová, J., et al. (1992). *Český jazykový atlas [Czech Linguistic Atlas]*, volume 1. Academia, Prague.

Balhar, J., Jančák, P., Bachmannová, J., Čižmárová, L., et al. (1997). *Český jazykový atlas [Czech Linguistic Atlas]*, volume 2. Academia, Prague.

Balhar, J., Bachmannová, J., Balátová, E., et al. (1999). *Český jazykový atlas [Czech Linguistic Atlas]*, volume 3. Academia, Prague.

Balhar, J., Bachmannová, J., Balátová, E., et al. (2002). *Český jazykový atlas [Czech Linguistic Atlas]*, volume 4. Academia, Prague.

Balhar, J., Bachmannová, J., Čižmárová, L., et al. (2005). *Český jazykový atlas [Czech Linguistic Atlas]*, volume 5. Academia, Prague.

Balhar, J., Bachmannová, J., Čižmárová, L., et al. (2011). *Český jazykový atlas – Dodatky [Czech Linguistic Atlas—Addenda]*. Academia, Prague.

Benešová, L., Křen, M., and Waclawičová, M. (2013). ORAL2013: reprezentativní korpus neformální mluvené češtiny [ORAL2013: A representative corpus of informal spoken Czech]. Ústav Českého národního korpusu FF UK, Praha.

[5] See for instance `http://www.arcgis.com/explorer/`.

[6] See `https://kontext.korpus.cz`.

Chambers, J. K. and Trudgill, P. (1998). *Dialectology*. CUP, 2nd edition.

Crowdy, S. (1993). Spoken corpus design and transcription. *Literary and Linguistic Computing*, 8(4):259–265.

Dialectological Commission of the Czech Academy of Sciences and Arts. (1951). *Pravidla pro vědecký přepis dialektických zápisů českých a slovenských [Rules for the Scientific Transcription of Dialectological Records of Czech and Slovak]*. Czech Academy of Sciences and Arts, Prague.

Feagin, C. (2002). Entering the community: Fieldwork. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 20–39. Blackwell Publishing, Malden, MA.

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

Hammer, L. (1985). *Prague Colloquial Czech: A case Study in Code Switching*. Ph.D. thesis, Bloomington.

Hutchby, I. and Wooffitt, R. (2009). *Conversation Analysis*. Polity Press, Cambridge (UK).

Kastner, Q. (1996). *Osidlování českého pohraničí od května 1945. Historická analýza doplněná kvalitativní sociologickou sondou. [Settling the Czech Borderlands from May 1945 Onwards. Historical Analysis and Qualitative Sociological Probe]*. Sociologický ústav AV ČR.

Kopřivová, M. and Waclawičová, M. (2005). Construction of spoken corpus based on the material from the language area of Bohemia. In Garabík, R., editor, *Computer Treatment of Slavic and East European Languages*, pages 137–40. Veda, Bratislava.

König, W. (2005). *dtv-Atlas Deutsche Sprache*. Deutscher Taschenbuch Verlag, München.

Mikolov, T., Oparin, I., Glembek, O., Burget, L., Karafiát, M., and Černocký, J. (2008). Použití mluvených korpusů ve vývoji systému pro rozpoznávání českých přednášek [Using spoken corpora for developing a system for transcribing lectures in Czech. In Kopřivová, M. and Waclawičová, M., editors, *Čeština v mluveném korpusu [Czech in Spoken Corpora]*, pages 161–6. Nakladatelství Lidové noviny, Prague.

Mitianoudis, N. (2004). *Audio Source Separation Using Independent Component Analysis*. Ph.D. thesis, Queen Mary, University of London.

Rychlý, P. (2007). Manatee/Bonito—A modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masaryk University.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category—ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Válková, L., Waclawičová, M., and Křen, M. (2012). Balanced data repository of spontaneous spoken Czech. In *Proceedings of LREC 2012*, pages 3345–9.

Waclawičová, M. and Křen, M. (2008). ORAL2008: New balanced corpus of spoken Czech. In *Proceedings of the International Conference "Corpus Linguistics—2008"*, pages 105–12, Saint Petersburg. Saint Petersburg University Press.

Wilson, J. (2010). *Moravians in Prague: A Sociolinguistic Study of Dialect Contact in the Czech Republic*. Peter Lang, Frankfurt am Main.

Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. (2009). Annis: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009, July 20-23*, Liverpool, UK.

Čermák, F. (1993). Spoken Czech. In Eckert, E., editor, *Varieties of Czech. Studies in Czech Sociolinguistics*, pages 27–41. Rodopi, Amsterdam/Atlanta.

Čermák, F. (2009). Spoken corpora design. Their constitutive parameters. *International Journal of Corpus Linguistics*, 14(1):113–123.