NEW SPOKEN CORPORA OF CZECH: ORTOFON AND DIALEKT

ZUZANA KOMRSKOVÁ – MARIE KOPŘIVOVÁ – DAVID LUKEŠ
– PETRA POUKAROVÁ – HANA GOLÁŇOVÁ[1]
[1] Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

KOMRSKOVÁ, Zuzana – KOPŘIVOVÁ, Marie – LUKEŠ, David – POUKAROVÁ, Petra – GOLÁŇOVÁ, Hana: New Spoken Corpora of Czech: ORTOFON and DIALEKT. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 219 – 228.

Abstract: The paper introduces the ORTOFON corpus of spontaneous spoken Czech and the DIALEKT corpus of Czech dialects, their design principles and practical solutions adopted during data collection.
Keywords: dialectology, lemmatization, spoken corpus, tagging, transcription

1 INTRODUCTION

This paper introduces new spoken corpora prepared by the Institute of the Czech National Corpus (ICNC). The process of collecting recordings for the ORTOFON and DIALEKT corpora started in 2012 and both have finally been published on June 2, 2017. Both corpora are lemmatized and morphologically tagged.

The ICNC has a long tradition in creating spoken corpora. The first corpus of spoken Czech was the Prague Spoken Corpus (PSC) [5] whose recordings span the years 1988–1992 and were made in the Prague area only. Its follower – the ORAL series corpora[1] – focused on spontaneous spoken conversations of family members or friends from different parts of the Czech Republic, in the course of their natural, usual interactions (e.g. at home during a meal, in a restaurant, in the street). Except for the last corpus in the ORAL series (ORAL2013 [3]), these corpora (namely PSC, ORAL2006 [15], and ORAL2008 [23]) have been published only as transcripts, without the corresponding sound recordings. By contrast, ORAL2013 provides access to the actual recordings aligned with a one-tier transcript.

While the new ORTOFON corpus follows this tradition as far as the manner of data collection is concerned, the DIALEKT corpus is a new project line which focuses on monological spoken language showcasing traditional dialects. Both new corpora are based on a multi-tier transcription setup.

2 THE ORTOFON CORPUS

This new spoken corpus of spontaneous everyday communication has been published on June 2, 2017, following several months of final data selection and revision. The data

[1] The ORAL series corpora were integrated into the ORAL corpus with 6 361707 tokens. This corpus is lemmatized and morphologically tagged in the same way as the ORTOFON and DIALEKT corpora. More at [18].

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0031

was collected during 2012–2017. In terms of linguistic annotation, it features lemmatization and morphological tagging (see section 4). The size of the final published corpus is 1 236 508 tokens. Like previous spoken corpora, the ORTOFON corpus is balanced with respect to several sociolinguistic categories.

The raw material consists of recordings of prototypical spoken language (Czech in our case) [7, p. 118], which is defined as informal and spontaneous conversations between people who know each other very well, situated in casual settings. The interactions take place in familiar environments (e.g. in private, among friends) and the situations are not experimentally induced. We only record adult speakers (18+ years old).

2.1 Metadata

Our external collaborators who record and transcribe the conversations were asked to provide a variety of information about each recording and each speaker. This information covers the two broad categories of "context-governed" and "demographic" details [4]. These enable the corpus user to restrict searches to specific types of extralinguistic context and to create subcorpora based on them. The goal is to capture as many of the factors which can possibly influence the conversation as possible.

The context-governed perspective covers general information about the recorded situation. There is a list of 12 pre-defined primary situation types, which distinguish the different possible settings in which the conversation could have taken place (for further details see [16], [17]). Another requirement is to enter the date, place, and corresponding geographical area of the recording location (the geographical areas are based on dialect areas which follow [1]). The collaborators are also asked to make a list of conversational topics and to fill them in. Apart from that, the relationship of speakers is indicated (one of partners, family, friends, acquaintances) and the total number of generations they represent (e.g. mother and daughter = two generations). There is also an assessment of the sound quality of the recording, which is useful for phonetic transcription. In the resulting corpus, the information related to the whole recording will be stored as per-document metadata.

The demographic perspective summarizes the speakers' characteristics; it is therefore mapped onto per-speaker metadata. In each recording, the speakers are numbered and cross-referenced with a speaker database. The database tracks the speakers' sociological characteristics, which include:

- gender
- age
- field and highest achieved level of education
- current and longest occupation
- childhood region and place of residence (until 15 years old), longest and current region and place of residence, and size of the corresponding administrative unit
- common speech defects.

2.2 Balancing the ORTOFON Corpus

The previous ORAL2008 and ORAL2013 corpora have been balanced according to three sociolinguistic variables: gender, age, and the highest achieved level of education.

Each variable was split into two levels (female × male, 18–34 years old × 35+ years old, non-tertiary × tertiary education) to avoid excessive fragmentation and to enable comparability with PSC. The balancing of the ORTOFON corpus is based on four sociolinguistic variables, namely the three previously mentioned ones and childhood region, which assumes ten dialect regions (see Fig. 1). The final corpus is trying to be representative (i.e. it includes speakers representing all possible combinations of the sociolinguistic variables, and as many different speakers as possible), and as balanced as possible (i.e. the proportions of all categories are roughly equal). Considering the target size of the corpus and the number of levels per the four variables, we get 1M / (2 × 2 × 2 × 10) = 12 500 tokens ideally for each combination, e.g. for female speakers 35+ y.o. with tertiary education from West Bohemia. We strove for a minimum of five different speakers per combination [9], which reduces the risk of a category being excessively tied to a single idiolect and maintains variability.[2]

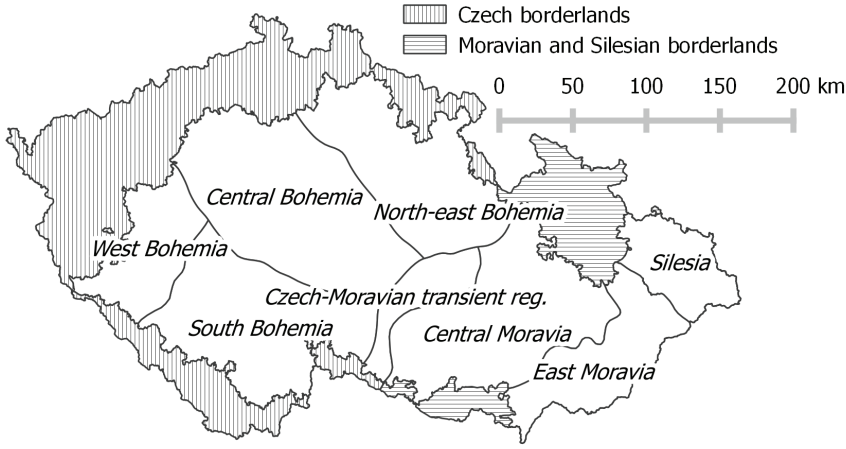

Fig. 1. Dialect regions in the ORTOFON corpus

The map shows all ten dialect regions. Their borders were determined according to several dialect studies (e.g. [14], [22]), so they have been slightly modified compared to ORAL2013.[3] While the previous ORAL series corpora only used the criterion of territory to a certain extent to make the data as representative as possible, ORTOFON treats the criterion of childhood territory on par with the other balancing variables.

2.3 Annotation Scheme

The main difference between the ORTOFON corpus and the ORAL series corpora is the multi-tier transcription. Every recording is transcribed using the ELAN[4] transcription software [21]. There are two main types of tiers (corresponding to

[2] More details at `http://wiki.korpus.cz/doku.php/cnk:ortofon`.

[3] The map is available at: `https://wiki.korpus.cz/lib/exe/detail.php/cnk:o13.png`.

[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/`.

orthographic and phonetic transcription) and each speaker in every recording gets their own instance of both of them, which means that overlaps may be transcribed in parallel on the respective orthographic (and phonetic) tiers of the overlapping speakers (there are always the whole words in overlaps, the overlapping speech is marked by square brackets []). Speakers' turns are segmented into sub-units of a maximum length of 25 tokens for ease of parallel revision.

The transcription workflow proceeds stepwise from a basic orthographic transcription with annotation of metalinguistic information, through revisions, and eventually to phonetic transcription.

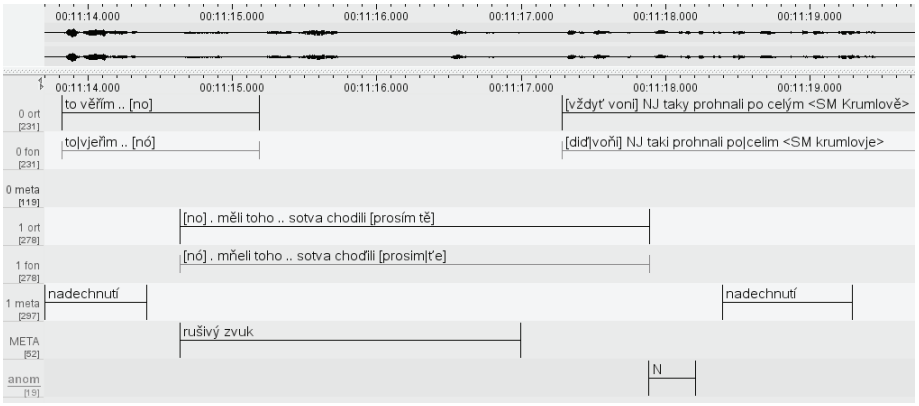

Fig. 2. Excerpt from a transcript for the ORTOFON corpus in the ELAN transcription program, showing the recording waveform at the top, with time-aligned orthographic, phonetic, and metalinguistic tiers for speaker 0 (0 ort, 0 fon, 0 meta) and speaker 1 (1 ort, 1 fon, 1 meta).

The multi-tier transcription shown in Fig. 2 illustrates the use of tiers: orthographic (ort), phonetic (fon), metalinguistic (meta, META), and anonymization (anom). The orthographic and phonetic tiers are reserved for speech transcription (see 2.3.1 and 2.3.2). Each speaker is further associated with their own metalinguistic tier (meta), which captures e.g. laughter or hiccups, i.e. paralinguistic sounds pertaining to a specific speaker, or pauses longer than two seconds. Additionally, there is another metalinguistic tier (META), only one instance per recording, whose purpose is to capture ambient sounds, e.g. phones ringing, dogs barking, or TV background noise. Both the meta and META tiers offer a list of pre-defined categories. Another layer (anom) is used for the anonymization of personal data, e.g. phone numbers, surnames, addresses. There is also a possibility to add another tier, the so-called JO tier, to capture the speech of a non-target speaker who disrupts the communication of target speakers, i.e. a waitress in a restaurant, or a child speaking to her mother. The anom and JO tiers are optional.

2.3.1 Orthographic Transcription

The starting point for annotation is the orthographic tier. It is optimized for a first quick transcription of the recording. Although the tier is named "orthographic", the transcription differs in some aspects from traditional written language. For instance, it captures dialectal features, e.g. variations in the endings for all types of conjugation

and declension. Conversely and unlike the ORAL series, it preserves the quantity of vowels according to standard Czech, all consonants in consonant clusters (e.g. *já vždycky vím* instead of pronounced *já dycky vim*), and full form of formally reduced variants of words (e.g. *myslím dostal šestnáct* instead of *sim dostal šesnáz*).[5] In case of two (or more) possible variants of transcription of the word, we choose only one of them (*citron/citrón* > *citron*; *osum/osm* > *osm*; *benzin/benzín* > *benzin* etc.).

A very important requirement is to ensure that the transcription procedure is homogeneous across different recordings, which already span over four years. For this purpose, we worked out a detailed manual for all our collaborators where they can find examples and general rules for transcription.[6] This manual has been continuously updated with additional examples gleaned from the material.

The most important phenomena captured on the orthographic tier include:

- v- or h-prothesis: ***v****okno*, ***h****ulica*
- regional variants of vocalic changes: *ml****ý****n - ml****ej****n - ml****é****n*, *l****ou****ka - l****ú****ka - l****ó****ka*
- regional declension variants: *s mal****ejma*** *nákład****ama*** (instead of *s malými náklady*)
- regional conjugation variants: *maj****í*** *- maj - maj****ú*** *- maj****ó*** (3-PL-*mít*), *chc****u*** *říct* (instead of *chci říct*)
- shortened forms of the 3rd pers. sg. past participle normally ending in -l: *moh*, *spad*, *řek*

Another specificity is pausal punctuation, used also in the ORAL2013 corpus. In the ORTOFON corpus, the term "pause" became more accurate, i.e. at least 120 ms of silence or other nonverbal sounds, e.g. breath, cough, laugh. However, pauses shorter than 120 ms may be annotated under the looser concept of "prosodic boundary", which also covers prosodic segmentation phenomena not implemented by an actual interruption of the flow of speech, like tempo changes and intonation cues. The transcription distinguishes three types of pauses with different symbols:

- . on the ort layer for prosodic boundaries (including pauses up to 120 ms);
- .. on the ort layer for pauses from 120 ms to 2 s;
- a separate segment annotated as *dlouhá pauza* (*long pause*) on the meta layer for pauses longer than 2 s.

The orthographic layer captures the verbal and near-verbal content of the interaction including unfinished words, false starts, hesitations, response sounds, and overlaps (for details on the particular symbols used, see [16], [17]).

Paralinguistic and situational comments are mainly captured on the meta and META layers, but some of them are also present on the orthographic tiers. This occurs when they are tightly coupled to a particular segment of speech: either because they could affect voice quality, e.g. laughter, yawning, loudness, or because they convey additional information, e.g. speech in foreign language, recitation, singing. The tokens uttered with that concomitant feature are signalized by angle brackets <>, e.g. *ty máš <SM nápady>*.[7]

[5] There is a list of formally reduced variants which have been lexicalized and thus transcribed, e.g. *čéče* (but *čoveče* is transcribed as a full *člověče*), *páč* (instead of *poněvadž*).

[6] Accessible at: `https://wiki.korpus.cz`.

[7] <*SM* ...> marks laughter.

2.3.2 Phonetic Transcription

The phonetic tier is an innovation compared to the ORAL series corpora. It has its own rules, which allow us to capture real pronunciation using a simplified phonetic transcription. Although it does not aim to capture all phonetic variation (e.g. the scale of vowel reduction), it still offers basic pointers concerning variability in spontaneous speech. Standard alphabet characters, extended with a small set of specialized symbols, are used instead of the International Phonetic Alphabet (for details on this decision see [16], [17]).

The phonetic layer is closely integrated with the orthographic layer. Some orthographic words are merged into prosodic words (or stress groups) on the phonetic tier, but the space between them is not simply removed. Instead, it is replaced with the pipe | symbol, so as to preserve information about the location of the orthographic boundary and, by extension, a one-to-one correspondence between the tokens on the two tiers. This allows search query constraints to target both tiers simultaneously, providing the users with more control over their search results.

The phonetic layer captures the following phenomena (in the example pairs, the first half corresponds to the ort layer and the second to fon):

- some non-phonemic distinctions, e.g. labiodental [ɱ] or velar [ŋ]: *prosím vás* → *prosi**ɱ**|vás*, *tenkrát* → *te**ŋ**krát*
- assimilations of voicing: *kup mi to* → *ku**b**|mi|to*, *tvoje* → *t**f**oe*
- assimilations of place of articulation: *hodně* → *ho**d'**ně* (see also examples under non-phonemic distinctions above)
- assimilations of manner of articulation: *od nás* → *o**n**|nás*
- shared phones, indicated via the underscore _ symbol: *dnes jsem se dobře vyspal* → *dne_|sem|se dobře vispal*
- epentheses and elisions: *zhasnout* → *zhas**t**nout*, *protože* → *bže*

3 THE DIALEKT CORPUS

This new corpus, published alongside ORTOFON, is our first attempt to build a collection of dialectal linguistic material compiled as a linguistic corpus. As far as we are aware, it is also the first dialectal corpus in the Czech Republic available through a user-friendly search interface, serving not only professional dialectologists but also the broader linguistics community, teachers and laypeople. Like the ORTOFON corpus, it is lemmatized and morphologically tagged.

The DIALEKT corpus differs from the ORTOFON in several characteristics. Firstly, it does not have a fixed size in tokens, it will be, hopefully, published regularly in versions with a growing amount of data.[8] The first version counts 128,289 tokens on the dialectological layer and 126,131 tokens on the orthographic layer. This is related to the second difference, that the corpus is not balanced, nor does it aim to be in the future. Thirdly, the material covers two broad stages of data collection: older data from the late 1950s up to the 1980s, which mostly comes from

[8] Creating a non-balanced, continuously growing version of the ORTOFON corpus, alongside the balanced one, is also under consideration.

the research effort which resulted in the Czech Linguistic Atlas [1], and new data since the 1990s [12]. This allows comparing the gradual loss of dialectal features in the respective dialectal regions. Additional differences concern the process of transcription (see 3.3).

3.1 Metadata

Due to the two stages of data collection, the metadata about speakers and the whole recording were adapted. The dialectal speakers had to fulfil certain criteria: they had to have spent the great majority of their life in a single rural area without moving to another dialectal region, they had to be over 60 years old and not university educated. There were no limitations as far as their occupation, but some speakers (teachers for example) usually adjust their speech or care much more about dialectal features, which influences their spontaneity. Speakers tied to traditional rural professions were therefore given preference, which goes hand in hand with an interest in dialectal lexis.

Regional classification is, in contrast to the ORTOFON corpus, more detailed. The ten dialectal regions, which are the same for both corpora, are divided into smaller sub-areas with a specific type of a particular dialect, and those can subdivided even further, according to the traditional three-level hierarchy for classifying dialects (*nářeční oblast* > *nářeční typ* > *nářeční úsek*). The metadata also show which region belongs to which territory of the Czech Republic, i. e. Bohemia, Moravia, Silesia, and if the type of residence was town or country. Further details about the metadata are available in [12].

3.2 Annotation Scheme

The recordings for the DIALEKT corpus are transcribed according to a similar procedure as the ORTOFON corpus, using the same tools. The types of tiers are the same with one exception: there is a dialectological layer instead of the phonetic one, and it is considered as the primary one (the primary layer for the ORTOFON corpus is the orthographic one).

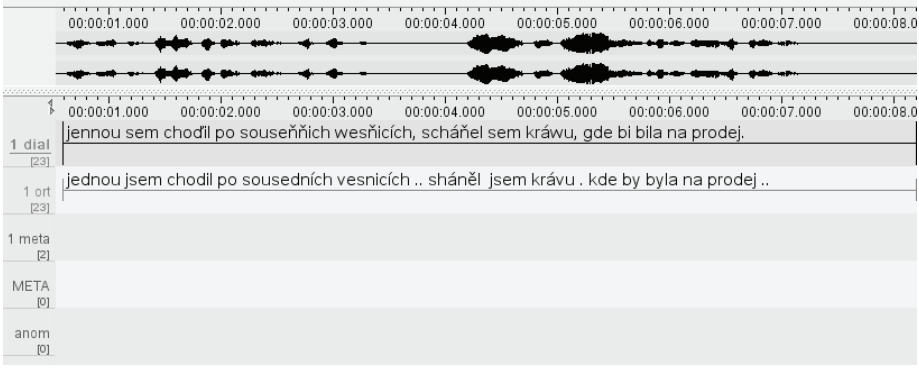

Fig. 3. Excerpt from a transcript for the DIALEKT corpus in the ELAN transcription program

3.2.1 Orthographic Transcription

The main reason for multi-tier transcription of dialectal data was comparability with other spoken corpora in the CNC, especially the ORTOFON corpus, the facilitation

of searching, and help for better lemmatization and tagging. But the richer variability in lexicon, morphology and phonology requires more aggressive standardization on the orthographic layer, which thus differs in some details from the corresponding one in the ORTOFON corpus.

The differences between the orthographic and dialectological tiers cover the following phenomena (the first word shows the transcription on the dialectological tier, the second its orthographic counterpart):

- v-prothesis is kept (***v**okno* > ***v**okno*), but h-prothesis is not (***h**erteple* > *erteple*)
- regional variants of vocalic changes are leveled on the orthographic tier: *k**ú**řilo s**a*** > *k**ou**řilo s**e***, *s**ej**tko* > *s**í**tko*
- regional variants of consonantic changes as well: *sva**r**eb* > *sva**t**eb*, *s**k**oval* > *s**ch**oval*, *kame**ň*** > *kám**en***

Other phenomena (e.g. vowel quantity, full form of consonant clusters and formally reduced variants, regional variants in declination and conjugation) are treated the same on the orthographic layers of both corpora.

3.2.2 Dialectological Transcription

The transcription rules for the dialectological layer are based on the usual conventions in the field of Czech dialectology.[9] This layer includes some specific symbols for dialectal vowels or consonants in order to capture the actual pronunciation, e.g. *vərch*, *býł*, *won*, *řezňičił*. In contrast, word boundaries are kept according the standard orthography and we use unrestricted syntactic punctuation, e.g. marking direct speech using quotes “”. Capital letters appear only at the beginning of proper names, like on the orthographic layer.

4 LEMMATIZATION AND TAGGING[10]

Even though the issue of lemmatization and tagging of spoken Czech has been discussed many times, practical attempts have been comparatively few, e.g. [9], [12], [13]. It is closely connected to the type of data (monologues, dialogues), and especially transcription rules, e.g. how the transcription is segmented, which type of punctuation is used, how much the transcript reflects real pronunciation etc. We decided to develop a pragmatically-minded custom solution based on existing and openly available tools, even though these are designed for written language.

The lemmatization and morphological tagging of both new spoken corpora were conducted according to the same process recently applied to the ORAL series [18]. We took the Czech morphological dictionary MorfFlex CZ [11] as a basis which has been manually and semi-automatic extended or cleaned according to the target register. The extensions refer mainly to register- and/or region-specific items, either full lexemes (lemmas *zbroják*, *škodárna*, *ikspéčka*) or inflectional variants (e.g. lemma *neděle* has two acc. sg. variants, *neděli* and *nedělu*), which were not

[9] We mostly follow the *Rules for the Scientific Transcription of Dialectological Records of Czech and Slovak* [8], but also take some inspiration from *Czech Dialectal Texts* [19] and the *Addenda to the Czech Linguistic Atlas* [5].

[10] For more information about lemmatization and tagging of the ORAL corpora see [18].

contained in the original morphological dictionary. Unsurprisingly, what makes lemmatization and tagging even remotely possible is the presence of an orthographic layer which is fairly close to standard language, at least in terms of transcribing the individual word forms.

5 CONCLUSION

Taken together, the ORTOFON and DIALEKT corpora allow users to research diachronic and diatopic variation in spoken Czech language through a convenient interface. Compared to previous spoken corpora built at the ICNC, they feature a more detailed annotation separated into several parallel layers accommodating speakers individually. The multi-tier transcription allows us to reserve one layer in both corpora for capturing pronunciation detail (be it from a phonetic – as in ORTOFON – or dialectological – as in DIALEKT – perspective), and another (called orthographic in both corpora) for general transcription. The orthographic layer serves as the basis for lemmatization and tagging of both spoken corpora.

This multi-tier transcription also presents challenges when indexing the corpora for querying with corpus tools which require a single authoritative tokenization of the text. A rigorous token-level alignment between the two tiers must be maintained at the transcription stage (as in the case of the ORTOFON corpus) or reconstructed (in the case of DIALEKT) in order to correctly link each token on the main layer with the corresponding token on the dependent layer.

A rich set of both context-dependent and demographic metadata provides additional perspectives on the collected material; especially the DIALEKT corpus provides useful information to researchers from related fields (sociologists, ethnographers, historians etc.). Both lines of data collection, as represented by the ORTOFON and DIALEKT corpora, will hopefully continue into the future.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

[1] Balhar, J. et al. (1992–2011). *Český jazykový atlas*. 6 sv. Academia, Praha.
[2] Balhar, J. et al. (2011). *Český jazykový atlas. Dodatky*. Academia, Praha.
[3] Benešová, L., Waclawičová, M., and Křen, M. (2013). ORAL2013: reprezentativní korpus neformální mluvené češtiny. ÚČNK FF UK, Praha. Accessible at: `http://korpus.cz`.
[4] Crowdy, S. (1993). Spoken Corpus Design and Transcription. *Literary and Linguistic Computing* 8(4):259–265.
[5] Čermák, F., Adamovičová, A., and Pešička, J. (2001). *PMK (Pražský mluvený korpus): přepisy nahrávek pražské mluvy z 90. let 20. století*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.

[6] Čermák, F. et al. (2007). *Frekvenční slovník mluvené češtiny*. Karolinum, Praha.
[7] Čermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics*, 14(1):113–123.
[8] Dialektologická komise České akademie věd a umění (1951). *Pravidla pro vědecký přepis dialektických zápisů českých a slovenských*. Česká akademie věd a umění, Praha.
[9] 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.
[10] Goláňová, H., Kopřivová, M., Lukeš, D., and Štěpán, M. (2015). Kartografické a geografické zpracování dat z mluvených korpusů. *Korpus – gramatika – axiologie*, 11:42–54.
[11] Hajič, J. and Hlaváčová, J. (2013). MorfFlex CZ. Univerzita Karlova v Praze, MFF, ÚFAL, Praha.
[12] Hlaváčková, D. (2001). Korpus mluvené češtiny z brněnského prostředí a jeho morfologické značkování. *Slovo a slovesnost*, 62(1):62–70.
[13] Hlaváčková, D. and Osolsobě, K. (2008). Morfologické značkování mluvených korpusů, zkušenosti a otevřené otázky. In Kopřivová, M. and Waclawičová, M., editors, *Čeština v mluveném korpusu*, pages 105–114, Nakladatelství Lidové noviny / Ústav Českého národního korpusu, Praha, Czech Republic.
[14] Kloferová, S. (2000). *Mluva v severomoravském pohraničí*. Masarykova univerzita, Brno.
[15] Kopřivová, M. and Waclawičová, M. (2006). *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.
[16] Kopřivová, M., Goláňová, H., Klimešová, P., and Lukeš, D. (2014). Mapping Diatopic and Diachronic Variation in Spoken Czech: the ORTOFON and DIALEKT Corpora. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 376–382, European Language Resources Association, Reykjavík, Iceland.
[17] Kopřivová, M., Goláňová, H., Klimešová, P., Komrsková, Z., and Lukeš, D. (2014). Multi-tier Transcription of Informal Spoken Czech: The ORTOFON Corpus Approach. In *Complex Visibles Out There*, pages 529–544, Univerzita Palackého v Olomouci, Olomouc, Czech Republic.
[18] Kopřivová, M., Komrsková, Z., Lukeš, D., and Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. *Korpus – gramatika – axiologie*, 15:47–67.
[19] Lamprecht, A. and Michálková, V., editors (1976). *České nářeční texty*. SPN, Praha.
[20] Lukeš, D., Klimešová, P., Komrsková, Z., and Kopřivová, M. (2015). Experimental tagging of the ORAL series corpora: Insights on using a stochastic tagger. In Král, P. and Matoušek, V., editors, *TSD 2015, LNAI 9302*, pages 342–350, Springer International Publishing.
[21] Sloetjes, H. and Wittenburg, P. (2008). Annotation by Category: ELAN and ISO DCR. In *LREC 2008: Sixth International Conference on Language Resources and Evaluation*, pages 816–820. Accessible at: `http://www.lrec-conf.org/proceedings/lrec2008/summaries/208.html`, retrieved 2017-07-31.
[22] Sochová, Z. (2001). *Lašská slovní zásoba*. Academia, Praha.
[23] Waclawičová, M., Kopřivová, M., Křen, M., and Válková, L. (2008). *ORAL2008: sociolingvisticky vyvážený korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.