sciendo
DOI 10.2478/jazcas-2019-0062

RELEVANT CRITERIA FOR SELECTION OF SPOKEN DATA: THEORY MEETS PRACTICE

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

KOPŘIVOVÁ, Marie – KOMRSKOVÁ, Zuzana – POUKAROVÁ, Petra – LUKEŠ, David: Relevant criteria for selection of spoken data: theory meets practice. Journal of Linguistics, 2019, Vol. 70, No 2, pp. 324 – 335.

Abstract: The present paper seeks to review relevant criteria used in classifying speech events (SEs) from the perspective of spoken corpus design. The primary goal is to survey the landscape of possible types of spoken language, so as to assess in which directions the coverage of spoken Czech offered by Czech National Corpus corpora can be expanded in the future. We approach the problem from both theoretical and practical points of view, examining what the theoretical literature has to say as well as approaches implemented in practice by existing spoken corpora of various languages. We then synthesize the obtained information into a pragmatically motivated set of SE classification criteria which does not aspire to be universal or definitive but aims to serve as a useful guiding principle and conceptual framework for understanding and promoting SE diversity when collecting spoken data.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 INTRODUCTION

Ever since spoken language corpora started appearing, their authors have been trying to include different types of spoken communication in them [1]. Deciding on the criteria for the composition of a corpus which aims to reflect spoken communication is a crucial part of the entire process, as building these corpora is very time consuming and costly ([2], [3], [4]). In addition, there is no generally accepted classification of spoken language which would be similar to the library classification of disciplines in written texts (Universal Decimal Classification). Spoken communication has a number of aspects which are difficult to project into one classification and even more difficult to implement within a corpus.

It is necessary to reconsider these criteria using literature and taking into account the practical solutions chosen by the authors of previous corpora. These considerations will help us to better target the collection of those types of spoken data which take an especially great deal of effort to collect.

There are many types of communication, some more specialized or less common than others, although in practice, they all lie along a continuum. To make the discussion manageable, we exclude communication with children or in general speakers who are still learning the language, as well as communication with animals or machines, and take into consideration only communication between adult humans proficient in the given language. We also restrict our notion of spoken language to utterances that are mostly formulated on-the-fly, as they are spoken. Therefore, we exclude written-to-be-spoken communication. We use the term speech event (SE) for a stretch of speech that takes place in a particular situation and under certain conditions (e.g. lecture – formal settings, prepared etc., conversation at dinner with friends – informal, spontaneous etc. [5]).

The article is structured as follows: the first chapter surveys theoretical approaches based on selected literature. The first part of the following section gives an overview of practical, actually implemented solutions on the example of selected corpora of spoken language, the second part contains a brief summary of currently available Czech spoken data. In the third chapter, we build on the ideas presented in the previous sections to present the factors that we believe are important for the collection of spoken data in Czech. In conclusion, we present and justify the current focus of spoken data collection at the Institute of the Czech National Corpus (ICNC).

2 OVERVIEW OF LITERATURE

The basic dichotomy in language, which is more or less apparent from the lowest (morphemic) to the highest (textual) level, distinguishes between spoken and written language. These constitute the two extremes of a scale (although this scale is of course a continuum), stereotypical/prototypical representatives with mutually exclusive features. Generally, this description appears in grammars where the terms "written" and "spoken" are implicitly connected with a use of language in particular situations which require fulfilling particular "norms". Accounts of written language usually concern the standard language (in the sense of prescriptive rules enforced in language). As for spoken language, it usually refers to spontaneous language used in informal settings among friends or family members etc. The term for this type of spoken language varies, e.g. common spoken language (e.g. [6], [7]), the language of everyday spoken dialogues [8], vernacular speech [9], [10, p. 233].[1]

The vagueness and inconsistency in terminology is also reflected in the inconsistency of the notion of the term "spoken Czech" itself. According to [11, p. 46], spoken Czech is understood in various ways in linguistic papers: as a "communica-

[1] [10] mentions the form of tales and dialect texts. For other possible terms used in the Czech context, see [11, p. 46]. They are all very descriptive, capturing the external conditions of the SE or its emotional setting (emotionality, expressive speech, intimate tone etc.). Another term from English-speaking studies is intimate discourse [12].

tion form / mode of being of language" where all spoken varieties are assigned, as a synonym for "SEs in standard Czech", typical for formal settings, or its meaning is reduced to dialects or Common Czech.

Even so, there are some conditions that are considered regarding the classification of SEs which we can generalize from (for more details, see [13]) and which could be helpful for describing the continuum between the two extreme, prototypical language forms: relationship between speaker and addressee, topic, shared context (not only regarding knowledge, but shared experience in general)[2], place and time[3] of SE, setting (official etc.), social status of speaker and addressee. These criteria are not completely orthogonal, i.e. specifying some of them might implicitly narrow down the possible values for others – e.g. given a specific topic and an official setting for an SE, we can reasonably expect preparedness and so on.

Not all criteria have to be taken into consideration, only some of them can be chosen for the classification of SEs, depending on the research topic – specifically, only criteria which correspond to the researcher's intention and aim can be taken into consideration, ultimately yielding not one universal classification, but various special purpose ones. As a consequence, the spectrum or continuum of SEs does not have to be defined exhaustively, but only selectively.

3 OVERVIEW OF SELECTED SPOKEN CORPORA

The *creation* of a spoken corpus is a challenge involving a number of smaller decisions on several levels. The design of the corpus should take into account the various dimensions underlying the variation that can be observed in language use. This chapter briefly summarizes the basic information about publicly available spoken corpora in six languages other than Czech (3.1) and Czech (3.2); the attention is paid to the types of SE gathered within the corpora.

3.1 Overview of selected non-Czech spoken corpora

3.1.1 Lancaster/IBM Spoken Corpus

This spoken corpus is the smallest and oldest in this overview. It contains 52,637 words of spoken British English and was released in 1987. The aim of the corpus was to collect a sizeable sample of that type of spoken English which is "suitable as a model for speech synthesis. This explains the relatively high proportion of prepared or

[2] [14] accentuates the relationship between the participants of an SE and the amount of shared context (lack of shared context requires adding "background information"; p. 40). In his book, attention is paid to five styles of English usage (intimate, casual, consultative, formal, frozen), which are situated on the scale of familiarity – formality.

[3] For example, [15, pp. 34–35] states four functions of spoken language. The most important difference is distinguishing between situational (commonly spoken) and non-situational SEs (marked as "secondary spoken"; lectures, expert training etc.; cf. [8, p. 189]).

semi-prepared speech produced by trained broadcasters" [16, p. 6]. The length of recordings was not limited. The corpus contains the following categories of SEs: commentaries, news broadcasts, religious broadcasts (= daily services), radio discussions, propaganda, university lectures, public lectures, magazine-style reporting, fiction and poetry readings, informal dialogues among friends. All categories except the last one were produced in a public, rather formal setting and for a public audience.

3.1.2 British National Corpus (BNC)

The design of the BNC has been perhaps the most impactful in terms of influencing subsequent spoken corpora. The original BNC, released in 1994, consists of two components, a 10-million-word spoken one and a 90-million-word written one. The newest version called BNC2014 follows the previous design while focusing on newer data, and as of March 2019, only the spoken part has been completed [17]. The spoken BNC2014 has 11.5 million words gathered from across the United Kingdom. In contrast to the spoken BNC1994, it contains only spoken interactions in informal settings, especially at home, which took place among friends and family members.

The spoken BNC1994 includes both spontaneous, informal interactions (in the so-called demographic part) and formal context-governed encounters in four broad categories of social context: education/providing information, business, institutional/public communication, and leisure. Each category within the context-governed part of the BNC1994 was limited in size (max. 200,000–300,000 words); the range of SEs was defined, but not fixed. "The overall aim was to achieve a balanced selection within each, taking into account such features as region, level, gender of speakers, and topic. Other features, such as purpose, were applied on the basis of post hoc judgements" [18]. Attention was also paid to the dichotomy of monologue (40% of each social-context group) and dialogue (60% of each social-context group).

3.1.3 Corpus Gesproken Nederlands (CGN)

The CGN initially had a carefully-structured design which however had to be revised for pragmatic reasons[4] [19]. In the overall design, the principal criterion was taken to be the socio-situational setting in which language is used, whereas communicative goal, medium, number of speakers participating, the relationship between speaker(s) and hearer(s), and the sociolinguistic characteristics of speakers (i.e. age, gender, region, socio-economic class) were seen as supplementary criteria.

The released version[5] has 8.916m words and more than half of this material (4.7m) comes from informal, spontaneous, face-to-face and telephone dialogues.

[4] These reasons are described in [20, p. 341] as follows: "… because of the time, financial, and legal constraints under which the project must operate, but also for practical reasons, it is impossible to include all possible types of speech and compromises are inevitable."

[5] For more details, see `http://lands.let.ru.nl/cgn/doc_English/topics/version_1.0/overview.htm`

The remaining part consists of the following SE categories: interviews with teachers, simulated business negotiations, broadcast interviews/discussions/debates, non-broadcast political discussions/debates/meetings, classroom lessons, lectures/seminars, broadcast live commentaries (e.g. in sports), broadcast news reports/reportage, broadcast news, broadcast commentaries/columns/reviews, ceremonious speeches/sermons, and read speech. The list shows that attention was paid to the distinction between monologues vs. dialogues, and broadcast vs. non-broadcast.

3.1.4 Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK)

The FOLK corpus was designed according to several principles inspired by previous spoken corpora, in particular the BNC1994. The aim was to gather "a maximally diverse range of verbal communication in private, institutional, and public settings" [21, p. 383]. The released version has 1.95m tokens and contains recordings collected within the FOLK project, as well as within other projects of the same institute, i.e. map tasks, biographical interviews. Data collection was relatively free as far as more detailed distinctions within the social-setting categories are concerned, although there was an effort to apply speaker-related sociolinguistic criteria (i.e. age, gender, region). For the complete list of SE categories, see [21].

3.1.5 Göteborg Spoken Language Corpus (GSLC)

The GSLC was created more opportunistically than the other corpora described here, which means without any prior corpus design. The main goal was to ensure the broadest possible range of different SEs [22]. The corpus consists of 1.42m words collected within 27 SE categories. Most of them could be found in any other spoken corpus (and cover the range of four social-setting categories in the BNC1994), but some are fairly unique (for example bus driver/passenger conversation, physical therapy, or quarrel). These latter were either produced in the workplace, e.g. in a factory, travel agency, hotel, shop, at the doctor's, or in a task-oriented experimental setting (the complete list is available in [22]). The categorization in the GSLC is mostly ad hoc and low-level, no effort was made to establish any higher-level categories according to e.g. number of speakers or situational context.

3.1.6 Slovenský hovorený korpus (SHK)

The SHK is a long-running project which regularly releases new and improved versions of spoken corpora of Slovak.[6] The newest version – s-hovor-6.0 – has 6.593m words. The SHK is a collection of recordings from various SEs within all sorts of social settings [23]. There are both dialogues and monologues with varying degrees of spontaneity and formality, e.g. spontaneous dialogues, lectures, sermons, broadcast discussions, oral-history narratives.

[6] For more details, see `https://korpus.sk/shk.html`

3.2 Publicly available spoken data of Czech

A disclaimer is in order first: the following discussion applies only to publicly available corpora mainly used for linguistic research, though we are well aware that there are other corpora (or more broadly, data sets) which may not be generally accessible and/or which serve other than linguistic purposes (e.g. speech recognition). With that in mind, most spoken Czech corpora to date have focused primarily on prototypical spoken language, which is defined as dialogic SEs within an informal private setting, among family members and friends [13]. Over the years, this niche has spawned parts of the PMK [24] and BMK corpora [25] the ORAL series corpora [26], and the ORTOFON corpus [27]. Unlike many other types of SEs, where pre-existing recordings can be harvested, this type of SE generally requires collecting data from scratch, i.e. fieldwork. The earliest available recordings of this sort date back to 1988; today, the ICNC continues collecting data in this tradition and aims to keep doing so for the foreseeable future.

Other types of SEs represented within public spoken Czech corpora include:

- the controlled interview, which is mainly used in dialect-oriented research and which usually takes place in a somewhat more formal setting; parts of the PMK and BMK corpora would fall into this category, as well as the DIALEKT corpus [28], [27]
- classroom interactions, in which dialogue is also prevalent, as gathered within the SCHOLA corpus of school communication, capturing entire lessons [29]
- broadcast TV programs, mainly debates and talk shows, collected in the DIALOG corpus [30].

This short overview suggests some areas that are not covered by the currently available roster of corpora, for instance monologues and non-broadcast SEs in a public setting.

3.3 Summary

This overview has shown different approaches to creating spoken corpora, from carefully planned to completely unplanned data collection. We emphasized those corpora that include the most diverse spoken data in terms of various characteristics of SEs. The overview of SEs included within these corpora serves as a source of inspiration in terms of the possible directions of expansion of spoken data collection in Czech.

4 CLASSIFICATION CRITERIA RELEVANT FOR SPOKEN CORPUS DESIGN

In order to inform current and future spoken data collection, we have chosen some aspects from the available theoretical descriptions and practical implementations of spoken data classification. In what follows, we foreground criteria which

can be derived directly from the situation in which the SE occurred, without having to ask the participants. Accordingly, some of the aspects mentioned above will tend to be downplayed, others may be amalgamated into a single criterion. Not all criteria apply to all SEs. One clear exception to the focus on situation-derivable criteria are socio-demographic characteristics, which participants typically need to state explicitly. However, their relevance to spoken corpus design is clear, so we mention them in a separate section for completeness' sake.

The criteria are divided into two broad categories: setting-related and participant-related. Each category consists of multiple subcategories. In order to describe them, we offer a simplified, often dichotomous account (cf. also [6]), but it goes without saying that the characteristics of real-life SEs are often far from black and white. When selecting and organizing the criteria, an effort was made to map them as closely as possible to the existing categories mentioned above, and to avoid ambiguous terms (e.g. 'spontaneous', which can mean either 'informal' or 'unprepared').

4.1 Setting-related criteria

4.1.1 Degree of officiality

This aspect takes into account the social role of the speaker. The term social role refers to the set of behaviors, rights, obligations, beliefs, and norms as conceptualized by people in a social situation [31]. This category distinguishes whether the speaker represents an institution[7] (e.g. the headmaster's opening speech at the beginning of the school year) or is entrusted with a ceremonial task to perform (e.g. a birthday toast). On the other side of the scale are situations where everybody can join the conversation at their own discretion.

4.1.2 Degree of publicness

This aspect is closely connected to the size of the audience and the relationship between them. It indicates whether the SE is public, accessible to everyone (e.g. a speech on a public square, a political debate on TV), or restricted to the members of the community within which the SE takes place (e.g. preaching, work training), or whether the SE only has one addressee and thus is private (e.g. conversation with a doctor, lawyer, friend).

4.1.3 Mediation of communication

There are situations where both participants are physically present in the same place (face-to-face), and situations where they are not and their communication is transmitted via a mediated channel (e.g. phone, Skype, live TV debate).

[7] The term *institution* is broadly understood as any generally practised pattern of behavior, regulated by a given culture, often associated with specific SEs [32].

4.1.4 Synchronicity

This criterion focuses on whether the communication takes place at the same time for both participants (e.g. face-to-face, via telephone) or if the time of speech production is distinct from the time of speech perception (e.g. pre-recorded material on TV or on the web).

4.2 Participant-related criteria

4.2.1 Number of (active) speakers

A prototypical monologue is a speech by one speaker who is informed in advance that the time for his/her speech will be reserved. Of course, in practice, verbal interaction with (one of) the addressee(s) can also be initiated, and monologues can also arise spontaneously from the situation, without being explicitly sanctioned. But the basic condition is that one speaker speaks and does not expect to be interrupted until s/he yields, whereas in a dialogue, the speaking role is shared by two or more participants and turn-taking is managed dynamically.

4.2.2 Degree of preparedness

Either the speaker knows about the purpose and topic of the SE and can therefore, at his/her discretion, make preparations (e.g. with prior research, presentation slides, written notes), or s/he does not know in advance that s/he will be speaking at a given moment and thus has to respond on-the-fly[8] (e.g. an opinion poll on the street, a private chat with friends).

4.2.3 Number of addressees

Instead of capturing this as a continuous numerical variable, it makes sense to partition the continuum into broad classes which capture some qualitative shifts. Sociologists distinguish the following levels: speech directed towards one person, a small group of up to 19 people, a large group up to 39 people or a large group of addressees (the public) [33, p. 68].

4.2.4 Degree of addressee activeness

We distinguish whether the addressee can claim speaking initiative and thus become an active speaker in his/her own right (as is typically the case in a dialogue), whether s/he is allowed to ask questions, or if s/he does not even speak or influence the other speaker at all (e.g. broadcast SEs).

4.2.5 Relationship between participants

4.2.5.1 Amount of shared background

This aspect captures the amount of common understanding of the wider context of the SE, which may be high e.g. for family members, long-time friends, but also

[8] This type of production is sometimes referred to as spontaneous, but as mentioned above, we refrain from using this term in this paper.

for professionals from the same field who are working on the same task. On the other side of the scale, there are SEs where participants share little background (private, professional or other) with respect to what is being spoken about.

4.2.5.2 Degree of familiarity

In this aspect, the closeness of participants is taken into account, i.e. how intimately acquainted they are with each other. To suggest the range possible here, consider e.g. the dynamic between family members vs. between applicant and recruiter in a job interview. Both types of situations can yield SEs which are private in the sense of 4.1.2 above, but they differ vastly in terms of their degree of familiarity.

4.2.5.3 Symmetry of social roles

Each SE is also influenced by the mutual social status of the participants in communication. This could be related to various factors, e.g. age or profession. The relationship between the social roles of both participants could be symmetric (e.g. conversation among friends of the same age), or asymmetric (e.g. conversation between boss and subordinate). When making social role symmetry judgments, it is important to realize that each participant plays many different social roles and to focus on the role(s) which is/are most saliently activated in the context of the given SE.

4.2.6 Socio-demographic characteristics

This category comprises the following kinds of items: gender, age, highest achieved level of education, region, profession, place of residence, size of settlement etc. It is a good idea to collect socio-demographic data which is as detailed as possible, and only later possibly bin the values into larger groups (e.g. age groups instead of looking at exact age). Without such binning, balancing or any kind of demographic representativeness (possibly even reflecting the demographic distribution in the population) may be an impossible goal to achieve.

5 CONCLUSION

We addressed the topic of classification of SEs, trying to summarize theoretical approaches and confront them with practical implementations in existing spoken corpora. After reviewing the theoretical literature and the composition of several spoken corpora, both Czech and non-Czech, we proceeded to sketch our own categorization, inspired by these sources. To make the criteria more specific, we attempted to exemplify them by suggesting extremes (for dichotomous categories) or points along the continuum. The resulting classification system can serve as an aid in identifying the types of SEs that are still missing from corpora of spoken Czech.

To wit, there is so far no Czech corpus that contains monological SEs from official settings like most of the spoken corpora described under 3.1 do. This is one of

the shortcomings that we aim to remedy in further data collection, focusing on this type of SE in various contexts, especially public, including professional lectures. Since for the time being, we would like to avoid material that has been extensively scripted and/or edited prior to broadcasting or publishing, we are not currently planning to collect monological SEs from mass media such as radio, television and web shows.

So far, spoken Czech corpora at the ICNC have been mainly focusing on only one type of SE, prototypical spoken language. While we would like to broaden our scope by including new SE types, many of the lessons learned in the past directly carry over: we can take advantage of previous experience and existing infrastructure. For instance, even with the new SE types, we can adhere to the same transcription process, using the same software and battle-tested transcription guidelines (possibly with minor adjustments where necessary or practical). This will streamline the entire process, as well as hopefully make it easier to search across different spoken corpora, and possibly even allow us to combine them into one super-representative corpus in the future.

ACKNOWLEDGMENTS

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] Svartvik, J. (ed.) (1990). The London-Lund Corpus of Spoken English: Description and Research. Lund Studies in English 82.

[2] Deppermann, A., and Hartung, M. (2012). Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des "Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim). In Felder, E., Müller, M., and Vogel, F. (eds). Korpuspragmatik, pages 414–450, Berlin, de Gruyter.

[3] Kopřivová, M. (2017). Mluvený korpus. In P. Karlík, M. Nekula, and J. Pleskalová (eds.), CzechEncy – Nový encyklopedický slovník češtiny.

[4] Gajdošová, K., and Šimková, M. (2018). Frekvenčný slovník hovorenej slovenčiny na báze Slovenského hovoreného korpusu. Bratislava, VEDA.

[5] Hirschová, M. (2017). Komunikační situace. In Karlík, P., Nekula, M., and Pleskalová, J. (eds.), CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: `https://www.czechency.org/slovnik/KOMUNIKAČNÍ SITUACE`.

[6] Chloupek, J. (1986). Dichotomie spisovnosti a nespisovnosti. Brno, Filozofická fakulta. Spisy univerzity J. E. Purkyně v Brně.

[7] Daneš, F. et al. (1997). Český jazyk na přelomu tisíciletí. Praha, Academia.

[8] Hoffmannová, J. et al. (2016). Stylistika mluvené a psané češtiny. Praha, Academia.
[9] Ervin-Tripp, S. M. (1964). An Analysis of the Interaction of Language, Topic and Listener. American Anthropologist 66, pages 86–102.
[10] Vachek, J. (1942). Psaný jazyk a pravopis. In Čtení o jazyce a poesii, pages 231–306.
[11] Hoffmannová, J. and Zeman, J. (2017). Výzkum syntaxe mluvené češtiny: inventarizace problémů, Slovo a slovesnost 78(1), pages 45–66.
[12] Clancy, B. (2015). Investigating Intimate Discourse: Exploring the spoken interaction of families, couples and friends. Routledge.
[13] Čermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. International Journal of Corpus Linguistics 14(1), pages 113–123.
[14] Joos, M. (1967). The five clocks. New York, Harcourt Brace & World.
[15] Chloupek, J. (1995). Sjednocující a rozrůžňující faktory v mluvené komunikaci. In K diferenciaci současného mluveného jazyka, pages 33–39, Ostrava, Repronis.
[16] Knowles, G., Taylor, L., and Williams, B. (1996). A Corpus of Formal British English Speech: The Lancaster/IBM Spoken English. Routledge, London & NY.
[17] Love, R., Dembry, C. Hardie A., Brezina, V., and McEnery, T. (2017). The Spoken BNC2014. Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, pages 319–344.
[18] Burnard, L. (ed.) (2000). The British National Corpus Users Reference Guide. Accessible at: http://www.natcorp.ox.ac.uk/docs/userManual/
[19] Oostdijk, N. (2002). The Design of the Spoken Dutch Corpus. In Peters, P., Collins, P., and Smith, A. (eds.), New Frontiers of Corpus Research. Amsterdam, pages 105–112.
[20] Oostdijk, N. et al. (2002). Experiences from the Spoken Dutch Corpus Project. Proceedings of the LREC 2002, pages 340–347.
[21] Schmidt, T. (2014). The Research and Teaching Corpus of Spoken German – FOLK. In Proceedings of the Ninth International conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland: European Language Resources Association (ELRA).
[22] Allwood, J. et al. (2003). Annotations and Tools for an Activity Based Spoken Language Corpus. In van Kuppevelt, Jan C.J., and Smith, R.W. (eds.), Current and New Directions in Discourse and Dialogue, pages 1–18, Springer.
[23] Šimková, M., Garabík, R., Karčová, A., and Gajdošová, K. (2008). Hovorený korpus slovenčiny. In M. Kopřivová, and M. Waclawičová: Čeština v mluveném korpusu, pages 227–233, Praha, NLN – ÚČNK.
[24] Čermák, F. et al. (2007). Frekvenční slovník mluvené češtiny. Praha, Karolinum.
[25] Hladká, Z. (2005). Zkušenosti s tvorbou korpusů češtiny v ÚČJ FF MU v Brně. In SPFFBU A 53, pages 115–124. Brno, Masarykova univerzita. Accessible at: http://hdl.handle.net/11222.digilib/101736
[26] Kopřivová, M., Lukeš, D., Komrsková, Z., and Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. In Korpus – Gramatika – Axiologie 15, pages 47–67.
[27] Komrsková, Z., Kopřivová, M., Lukeš, D., Poukarová, P., and Goláňová, H. (2017). New Spoken Corpora of Czech: ORTOFON and DIALEKT. Jazykovedný časopis, 68(2), pages 219–228.
[28] Goláňová, H. (2015): A new dialect corpus: DIALEKT. In Gajdošová, K., and Žáková, A. (eds.): Proceedings of the Eight International Conference Slovko 2015 (Natural Language Processing, Corpus Linguistics, Lexicography), pages 36–44. Lüdenscheid, RAM-Verlag.

[29] Šebesta, K. (2010): Korpusy češtiny a osvojování jazyka. Studie z aplikované lingvistiky, 2, pages 11–33. Accessible at: https://studiezaplikovanelingvistiky.ff.cuni.cz/wp-content/uploads/sites/19/2016/03/karel_sebesta_11-33.pdf
[30] Čmejrková, S., Jílková, L., and Kaderka, P. (2004). Mluvená čeština v televizních debatách: korpus DIALOG. Slovo a slovesnost, 65, pages 243–269.
[31] Vláčil, J. (2017). Role. In Z. R. Nešpor, editor, Sociologická encyklopedie. Praha, Sociologický ústav AV ČR, v.v.i. Accessible at: https://encyklopedie.soc.cas.cz/w/Role
[32] Keller, J. – Vláčil, J. (2017). Instituce. In Z. R. Nešpor (ed.), Sociologická encyklopedie. Praha, Sociologický ústav AV ČR, v.v.i. Accessible at: https://encyklopedie.soc.cas.cz/w/Instituce
[33] Novotná, E. (2010). Sociologie sociálních skupin. Praha, Grada.