The Nijmegen Corpus of Casual Czech

Mirjam Ernestus[1,2], Lucie Kočková-Amortová[1], Petr Pollak[3]
[1] Centre for Language Studies, Radboud University Nijmegen
[2] Max Planck Institute for Psycholinguistics
[3] Faculty of Electrical Engineering, Czech Technical University in Prague
[1,2]Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[3]Technicka 2, 166 27 Praha 6, Czech Republic
E-mail: m.ernestus@let.ru.nl, lucik.kocka@seznam.cz, pollak@fel.cvut.cz

Abstract

This article introduces a new speech corpus, the Nijmegen Corpus of Casual Czech (NCCCz), which contains more than 30 hours of high-quality recordings of casual conversations in Common Czech, among ten groups of three male and ten groups of three female friends. All speakers were native speakers of Czech, raised in Prague or in the region of Central Bohemia, and were between 19 and 26 years old. Every group of speakers consisted of one confederate, who was instructed to keep the conversations lively, and two speakers naive to the purposes of the recordings. The naive speakers were engaged in conversations for approximately 90 minutes, while the confederate joined them for approximately the last 72 minutes. The corpus was orthographically annotated by experienced transcribers and this orthographic transcription was aligned with the speech signal. In addition, the conversations were videotaped. This corpus can form the basis for all types of research on casual conversations in Czech, including phonetic research and research on how to improve automatic speech recognition. The corpus will be freely available.

Keywords: corpus of informal conversations, Common Czech, orthographic transcriptions

1. A new corpus of Common Czech

In the past decades, researchers have become increasingly interested in language spoken in naturally occurring social interaction (e.g. Local, 2003) and have started to investigate the characteristics of spontaneous and casual speech in languages such as German, Dutch, and English (e.g. Kohler, 1990; Ernestus, 2000; Johnson, 2004). Similarly, during the last fifteen years, several Czech linguists have shifted their attention from the written literary form, which had been investigated and well documented in the past, towards spoken Czech. These linguists have, among others, drawn attention to the fact that especially the variety known as Common Czech needs to be investigated, since it is now the prime spoken variety for most Czech speakers in informal communication (e.g. Čermák, 1997).

So far, a number of research institutes have created Czech corpora for many research goals, including Czech variants of *SpeeCon* and *SpeechDat* (Černocký & Pollák, 1999; Siemund et al., 2000), *Czech Radio Broadcast News* (Psutka et al., 2001) and the *Pražský Fonetický Korpus* (*Prague Phonetic Corpus*, Volín et al., 2008). These corpora typically contain read speech or speech produced during formal interviews. To our knowledge, no existing corpus of Czech contains high-quality recordings of naturally occurring interactions which are suitable for detailed phonetic research on casual speech.

Our paper introduces a new corpus, the Nijmegen Corpus of Casual Czech (NCCCz), which was created to fill this gap[1]. We also present some simple analyses based on the orthographic transcriptions, showing the degree of casualness of the recorded speech and the research possibilities of the corpus.

This paper is organized as follows. Section 2 shortly discusses the language situation in the Czech Republic and the language variety called Common Czech. Section 3 presents the NCCCz, while Section 4 presents initial research based on the orthographic transcriptions of this corpus. We summarize and conclude the paper in Section 5.

2. Common Czech

In the Czech Republic, Literary Czech is the language taught at school and the language that is used in formal writing and formal oral interactions (e.g. Čechová, 2000). However, due to the increasing prestige of informal language, and due to the influence of spoken mass media, Literary Czech now appears to be in the process of being replaced by Common Czech: Common Czech is now also used in formal interactions in the whole Czech Republic (e.g. Čmejrková et al., 2004).

The majority of Czech linguists define Common Czech as having those properties of the informal language that do not occur in Literary Czech. As a consequence, the spoken informal language is assumed to be a mixture of Literary Czech and Common Czech (e.g. Kopřivová & Waclawičová, 2008; Čechová, 2000) and Czech speakers would therefore be continuously code-switching. In line with Čermák (1996), in this paper, we consider Common Czech as being the complete informal spoken language, including properties that do and do not occur in Literary Czech.

The most important differences between Literary and Common Czech are found in morphology. First, the two language variants differ in a number of grammatical morphemes. Common Czech

[1] See http://mirjamernestus.ruhosting.nl/Ernestus/NCCCz for information on how to obtain a copy of the corpus.

• contains the unified instrumental ending *–ma*, instead of the standard ending *–mi*, for nouns, adjectives, pronouns, and numerals in the plural (e.g. *spojenej-ma sila-ma* '(with) joined forces');
• has lost gender distinction in the plural adjective and pronoun declensions (e.g. *ty mal-ý psi*/*lesy*/*boty*/*kola* 'these small dogs/forests/shoes/bikes' instead of *mal-í psi*, *mal-é lesy*, *mal-é boty*, *mal-á kola*);
• has different forms for the auxiliary verb 'to be' in the first person conditional (e.g. *bysem* instead of *bych* '(I) would').

Second, some morphemes show different pronunciations in Literary and Common Czech (e.g. Krčmová, 2000). In Common Czech:

• the morpheme [i:] is pronounced as [ej] in regular endings of the third person plural present tense form of some verbs (e.g., [musej] '(they) must' instead of [musi:]) and in the declension of consonant-final adjectives (e.g. [mladej] 'young' instead of [mladi:]);
• the morpheme [i:] is pronounced as [i] in the declension of adjectives (e.g., [mladim] 'young' instead of [mladi:m]);
• the morpheme [ɛ:] is pronounced as [i] in the declension of consonant-final adjectives (e.g., [mladim] 'young' instead of [mladɛ:m]);
• the morpheme [ɛ:] is pronounced as [i:] in the declension of consonant-final adjectives (e.g., [mladi:ɦo] '(without) young' instead of [mladɛ:ɦo]).

Even though Common Czech often appears in written form (especially in private letters and in contemporary Czech prose reflecting spontaneous speech), there is neither an official grammar nor a dictionary, nor is there an official spelling of Common Czech (e.g. Čermák, 1996). This raises the question how to transcribe Common Czech corpora orthographically (see Section 3.2).

3. The Nijmegen Corpus of Casual Czech

The Nijmegen Corpus of Casual Czech (NCCCz) consists of 30 hours of informal conversations recorded in November 2008 at Charles University in Prague. All 60 speakers (30 males and 30 females) have a similar geographical and educational background which allows researchers to study inter-speaker variation. The corpus also allows for research on within-speaker variability, since every group of three speakers was recorded for approximately 90 minutes. The conversations were videotaped as well, providing information on facial expressions and body movements (e.g. gestures). All these characteristics make the NCCCz a unique contribution to the state-of-the-art of corpora for casual Czech. In the following sections, we describe the data collection and the orthographic transcription procedures for the corpus in detail and provide an overview of the speech in the corpus.

3.1 Data collection

For the NCCCz, we followed the same procedure that was used for the collection of the Nijmegen Corpora of Casual French and Spanish. This procedure elicits natural and casual speech (Torreira, Adda-Decker & Ernestus, 2010; Torreira & Ernestus, 2012).

Twenty speakers acted as confederates and were asked to find two friends of the same sex (henceforth the naive speakers) willing to participate in recordings of natural conversations. In total, we recorded 20 sessions involving 60 speakers (10 groups of male and 10 groups of female speakers). Forty-nine speakers came from Prague, the remaining eleven from the central part of Bohemia. All were native speakers of Czech, aged between 19 and 26 (average 20.6), and had successfully finished high school. None had received any phonetic or dramatic education. Except for three speakers, who had full time jobs, all were university or college students. None of them reported any speech or hearing disorders.

Each session was recorded in a soundproof booth with an approximate size of 2.5 m by 3.5 m. The speakers sat on chairs around a table. All speakers were recorded on separate audio channels by means of two Edirol R-09 solid-state stereo recorders (one for the naive speakers, placed outside of the booth; one for the confederate, placed under the table inside the booth), three Samson QV head-mounted unidirectional microphones and two stereo microphone preamplifiers. The microphones, which were placed at an average distance of five cm from the left corner of the speakers' lips, were hypercardioid, which minimized cross-talk between speakers. The sampling rate was 44.1 kHz and quantization was set to 32 bits. Each session was also videotaped (Canon XM2 Mini-DV video camera), without the naive speakers being aware of this. Since the camera could only record two speakers, we decided not to video tape the confederate.

Our recording procedure included a preparative part and the recording itself, which was divided into three different parts. For the preparative part, the confederate arrived thirty minutes earlier than the two naive speakers at Charles University, without telling them about this meeting. The second author (henceforth experimenter) informed the confederate that the naive speakers would be filmed, and asked the confederate to take the chair outside the recording range of the camera. The confederate was also provided with instructions for each part of the recordings (see below).

At the beginning of the first part of the recording, the experimenter told the speakers to turn off their cell-phones and left the booth. The confederate then pretended to have received an important message or phone call that had to be answered immediately and left the booth as well. The two naive speakers were left alone without information about whether they were already being recorded.

Depending on the liveliness of the conversations between the two naive speakers, after a period of 11 to 24 minutes (18' 29" on average) from the beginning of the recording, the experimenter asked the confederate to return to the booth. This marked the beginning of the second part of the recording, which consisted of free conversation among the three speakers. As we were above all interested in the

speech produced by the naive speakers, confederates were asked to participate in the conversations only when necessary, to keep them lively. Various topics were addressed during the conversations, including school, relationships, common hobbies, and stories about all sorts of meetings.

The third part of the recordings started after a period of 53 to 68 minutes (63' 53'' on average) from the beginning of the recordings. The experimenter entered the room with a list of questions about political and social issues and the speakers were asked to discuss at least four issues from the list and to negotiate a common opinion for each question.

At the end of the recording session, we revealed our procedure to the naïve speakers. All speakers were paid and voluntarily signed a consent form giving researchers permission to use the audio and video recordings for scientific purposes. The speakers were given the opportunity to formulate restrictions regarding the use of their recordings, but none of them did so.

3.2 Orthographic transcription

The corpus was orthographically transcribed by native speakers of Czech, who used the TRANSCRIBER software (Barras et al., 2001). We provided the transcribers with detailed transcription guidelines based on those developed by LIMSI (Gauvain et al., 2002). For each speaker, we created mono-channel audio streams and separate annotation files. If the transcribers needed the speech of the other speakers in order to better understand the context of the speaker's words, they thus had to listen to the signals of the other speakers.

Both the speech and non-speech events, such as laughter, were manually segmented into small chunks. These chunks were orthographically transcribed, including markers for typical speaker noises, such as breath, laughter, cough, clicks and filled pauses. In total, the transcribers annotated 68,426 chunks (including standalone non-word events such as noise or laughter) with an average duration of 2.37 seconds. Stretches of silence within chunks were maximally 500 ms.

As mentioned in Section 2, there is no official spelling for Common Czech. Nevertheless, we decided not to transcribe the words in Literary Czech, especially as the inflectional word forms of Common and Literary Czech differ substantially. Moreover, Literary Czech forms cannot reflect some connotations specific to casual speech (e.g. vulgarity).

In contrast to the transcription conventions for other corpora of spoken Czech (e.g. those mentioned in section 1), we opted for a purely orthographic transcription of the words in our corpus and restricted the registration of pronunciation variation to the minimum. That is, we used only one orthographic form for every word, even if the different tokens of a word showed clear pronunciation variation. Thus, the transcribers were asked not to register prothetic [v-], the shortening or lengthening of vowels (including the difference between standard /i:/ and Common Czech /i/), or elisions (as for [ɲɛjaki:] produced as [ɲa:ki:] 'some', or for [nɛsl] produced as [nɛs] 'he carried'). There were two reasons for this. Firstly, the representation of each word type with just one orthographic word form is convenient for searching the corpus. Secondly, providing detailed phonetic (or quasi-phonetic) transcriptions is extremely time-consuming, as well as subjective and error prone.

All digits were transcribed as full orthographic words. Broken words were marked with brackets at the end of the pronounced part (e.g., "pro()") and the dollar sign preceded spelled words and letters (e.g., the letter *k* pronounced as [ka:] was transcribed as "$ká"). Moreover, all filled pauses were annotated with a single symbol and the registration of phonetic variation in interjections was restricted to the minimum (e.g., "br" and "brrrr" were both transcribed as "br"). Furthermore, we used only three punctuation marks (".", "?" and ","). Capital letters were only kept in proper nouns and acronyms.

The quality of an orthographic transcription can best be checked by comparing it to the independent transcription of an expert transcriber. Therefore, supervising experts (the last two co-authors of this paper) checked random parts of the transcriptions of all sessions. In case of insufficient quality, the transcriber was asked for revision, which was then checked again. The expert transcribers checked the first transcriptions of each new transcriber with special care in order to ensure uniform transcriptions as much as possible. We also checked the transcriptions against an approved lexicon of Czech, the Czech LC-StarII lexicon (Pollák et al., 2008). We added several Czech words that occurred in our corpus to this lexicon.

3.3 Speech in the corpus

The corpus comprises over 30 hours of recorded conversations. Table 1 shows the different types of speech and non-speech in the corpus (both in total and averaged over recording sessions). Non-speech includes silence, laughter and other types of noise produced by the speakers. The orthographic transcriptions contain 361,977 word tokens.

Two characteristics of the conversations indicate that these were lively and contained highly spontaneous, casual speech. First, the corpus contains relatively few stretches of silence (less than 7% of the whole corpus, approximately 7 minutes in total per session) or non-speech sounds other than laughter. Second, more than 20% of the speech in the NCCCz is produced in overlap.

Table 2 lists, per thousand words, the number of filled pauses, broken words, unintelligible words, and response vocalizations. These relatively high numbers also suggest that the corpus contains natural informal conversations.

4. Initial research based on the NCCCz

In this section, we discuss some phenomena which reflect the casual speech style in the corpus and for which the data can easily be extracted from the orthographic transcriptions. Causal speech may deviate from formal

	Total	Average	Max	Min	Total %
Speech	27h 50' 39"	1h 23' 32"	1h 27' 44"	1h 16' 18"	92.2
Overlapping speech	6h 52' 40"	20' 38"	36' 34"	10' 15"	22.8
Non-overlapping speech	20h 57' 59"	1h 02' 54"	1h 10' 45"	51' 11"	69.4
Non-speech with laughter	2h 21' 34"	7' 05"	13' 54"	3' 08"	7.8
Non speech without laughter	2h 05' 06"	6' 18"	6' 18"	3' 04"	6.9
Total – All Recordings	30h 12' 13"	1h 30' 37"	1h 32' 13"	1h 29' 57"	*100.0*

Table 1: Total duration (Total) and percentages (Total %) of Speech (Overlapping and Non-overlapping) and Non-speech (with and without laughter) in the NCCCz, along with average, minimum (Min) and maximum (Max) duration per session.

	NCCCz
Filled pauses	12.06
Broken words	11.45
Unintelligible speech	26.19
Response vocalizations	10.83

Table 2: Frequencies (normalized per thousand words) of Filled pauses, Broken words, stretches marked as Unintelligible speech, and Response vocalizations in the NCCCz.

speech at various levels (e.g. word choice, prosody, and frequencies of the different syntactic structures, including incomplete sentences). We focused on word choice since word choice can easily be investigated on the basis of the available orthographic transcriptions.

The analyses of these phenomena are based on all the speech in the corpus, as we found no differences between the naive speakers and the confederates. We also checked for differences between male and female speakers and we mention these differences if they were statistically significant.

4.1 Expressivity

4.1.1 Swear words

The presence of swear words is a strong indicator of the informal nature of speech. We therefore expected swear words to be frequent in the NCCCz. Whereas there is a Czech dictionary of swear words based on written texts (Ouředník, 2005), swear words in spoken language have received hardly any attention. In order to obtain a list of the most common swear words in spoken Czech, we asked twenty native speakers of Czech to provide a list of the ten most common Common Czech swear words. We grouped word forms with the same stem together and removed those stems that appeared in the respondents' lists only once. In this way, we obtained a list of 27 swear stems.

Two speakers used swear words with these stems very frequently (28 and 36 tokens per thousand word tokens), especially the word *vole* 'you sod' (24 and 32 tokens per thousand word tokens). They appear to use *vole* as a kind of filler word. The other speakers used swear words less frequently (on average twice per thousand word tokens). They most frequently used v*ůl*, *blb*, *prdel*, and *debi* (1.66, 0.83, 0.17 and 0.12 tokens per thousand word tokens, respectively). This high frequency of swear words confirms our characterization of the corpus as containing casual speech.

4.1.2 Diminutives

In Czech, normal (first-grade) diminutives can be converted into second-grade diminutives, which express intensification of the diminutiveness (e.g. *strom* 'tree' - *strom-ek* 'small tree' - *strom-eč-ek* 'very small tree') and very often also have affective connotations. Second-grade diminutives can easily be extracted from orthographic transcriptions because of their specific suffixes. Nevertheless, a further manual check is necessary in order to filter out non-diminutives ending in grapheme sequences identical to those of diminutive suffixes (e.g., *řidič-ky* 'drivers' versus *holč-ičky* 'very small girls').

The NCCCz contains 476 tokens of second-grade diminutives. Most speakers produced less than two second-grade diminutives per thousand word tokens, while one speaker produced more than five per thousand word tokens. A one-way ANOVA indicated a strong effect of gender ($F(1,58) = 14.98$, $p < 0.001$): Women produced second-grade diminutives more often than men (women: 304 tokens in total; men: 174 tokens in total), in line with the stereotype of women's speech being more affective.

The three most frequent second-grade diminutives are the inflectional forms of *babička* 'grandma', *maminka* 'mum' and *tatínek* 'daddy' (0.17, 0.09 and 0.06 tokens per thousand word tokens, respectively). These names of family members form approximately one quarter of all second-grade diminutives in the corpus. Interestingly, a considerable number of diminutives are *hapax legomena*: of the 164 stems used in second-grade diminutives, 96 stems appear only once in this type of diminutive. This

reflects the high productivity of second-grade diminutive formation in Czech. In addition, this high frequency is another indication of the casual speech style in the NCCCz.

4.2 Auxiliary verb *být* 'to be'

The use of the Czech auxiliary verb *být* 'to be' appears to be sensitive to speech style and this verb may therefore form another marker of the casual speech style of the NCCCz. We investigated the presence versus absence of this verb for the past tense and its form in the conditional.

4.2.1 Absence of *jsem* in the past tense

In Czech, the past tense is expressed by means of the past participle in the proper gender form (ending in *-l* for male speakers, ending in *-la* for female speakers) and the present tense form of the auxiliary verb *být* 'to be'. Whereas in Literary Czech, the auxiliary verb is always present for the first person singular, in Common Czech, the auxiliary verb may be absent (e.g. Bělič, 1972; Grepl et al., 1996). The past participle is then preceded (not necessarily immediately) by the personal pronoun *já* 'I' (e.g. *Já by-la* 'I been' versus standard *Já jsem by-la* 'I have been').

We investigated how often the past participle occurred without the auxiliary in the NCCCz by extracting all first person singular past participles. We split the corpus into two by the speaker's gender and extracted all words ending in *–l* from the male corpus and ending in *–la* from the female corpus. From these words, we automatically selected those that were preceded by *já* in the same sentence. Finally, we manually checked this list for words that were no past participles.

We fitted a linear mixed-effect regression model with the presence versus absence of *jsem* as a binomial dependent variable, with the speaker's gender and role (naive speaker versus confederate) and the overall frequency of the past participle as fixed effects, and speaker and stem of the past participle as crossed random effects. The results only showed a main effect of gender (β = -0.7373, z-value = 2.259, $p < 0.05$): *Jsem* is more often absent in men's speech (in 17% of the cases) than in women's (11%), which is as expected since men tend to reduce more than women do (e.g. Byrd, 1994). The effect remained significant after removal of the verb *myslet* 'to think', for which *jsem* is most often absent (15%). This shows that the gender effect is not driven by just this one verb. Moreover, these percentages show that our corpus contains informal speech.

4.2.2 Non-canonical forms expressing the conditional

Čmejrková (2005) pointed out that the use of canonical and non-canonical forms for the first person conditional of the auxiliary verb *být* 'to be' (canonical: *bych* '(I) would', *bychom* '(we) would', non-canonical: *bysem* '(I) would', *bysme* '(we) would') depends on language mode (written versus spoken) as well as on the style of a text. She also mentioned that the use of the canonical / non-canonical form is speaker dependent.

The NCCCz also shows substantial differences among speakers: The use of non-canonical forms ranges among the speakers in the NCCCz from 0% to 50% (mean: 13%), with 19 speakers (of the 60) not using any non-canonical form at all. We investigated when the auxiliary was canonical or non-canonical by means of a linear mixed-effect regression model with speaker as random variable and, as fixed effects, the speaker's gender and role, the grammatical number of the auxiliary (singular versus plural), and whether it carried a prefix and if so which (no prefix, prefix *a*, prefix *kdy*). This model showed that for most, but not for all speakers, non-canonical forms were more frequent for the plural than for the singular forms, and more frequent for morphologically simple than for prefixed verbs (i.e. the best model contained by speaker random slopes for grammatical number, $\chi^2 = 507.42$, df = 2, $p < 0.0001$, and for prefix, $\chi^2 = 172.67$, df = 3, $p < 0.0001$).

5 Summary and Conclusions

This article describes the Nijmegen Corpus of Casual Czech (NCCCz), which contains over 30 hours of Common Czech produced in informal conversations by 20 groups of three young adults. The corpus contains a large amount of speech for every speaker and therefore provides researchers with sufficient material for the study of inter- and intra-speaker variability. The quality of the speech recordings is high, allowing detailed acoustic research.

Our analyses of the orthographic transcriptions show that the speakers used many swear words and many second-grade diminutives, that they often deleted forms of *jsem* in the past tense, and that most speakers used non-canonical forms for the auxiliary expressing the conditional. These findings support our assumption that the corpus contains highly casual speech. The analyses also reveal that female speakers used second-grade diminutives more often than male speakers and that male speakers more often deleted forms of *jsem* in the past tense. Our findings therefore confirm the naturalness of the casual Czech speech in the corpus.

In conclusion, we believe that our Nijmegen Corpus of Casual Czech is a valuable source of information on casual Common Czech. We hope that many researchers will use the corpus for their study of all types of phenomena of Czech conversations.

6. Acknowledgements

Our thanks to the staff at the Phonetic Institute at Charles University in Prague for their help during the recordings of the corpus in Prague. Our special thanks to Lou Boves for valuable discussions. This work was funded by a European Young Investigator Award given to the first author. In addition, it was supported by two Czech grants

from GACR 102/08/0707 and CTU SGS 14/191/OHK3/3T/13.

7. References

Barras, C. ; Geoffrois, E. ; Wu, Z. and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2), pp. 5-22.

Bělič, J. (1972). *Nástin české dialektologie.* Prague: Státní pedagogické nakladatelství.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, 15(1-2), pp. 39-54.

Čechová, M. et al. (2000). *Čeština - řeč a jazyk*. Prague: ISV.

Čermák, F. (1996). Obecná a spisovná čeština: Poměr, funkce, a metodologie. In R. Šrámek (Ed.), *Spisovnost a nespisovnost dnes.* Brno: Masaryk University, pp. 14-18,

Čermák, F. (1997). Obecná čeština: je součástí české diglosie? *Jazykovědné, aktuality*, 34(3-4), pp. 34-43.

Černocký, J., Pollák, P. (1999). Specification of speech database interchange format. *SpeechDat-East Technical Report ED1.3*.

Čmejrková, S. ; Jílková, L. and Kaderka, P. (2004). Mluvená čeština v televizních debatách: korpus DIALOG. *Slovo a slovesnost*, 65(4), pp. 243-269.

Čmejrková, S. (2005). Bychom, nebo bysme? *Naše řeč*, 88(1), pp.18–36.

Ernestus, M. (2000). *Voice assimilation and segment reduction in Dutch: A corpus-based study of the phonology-phonetics interface*. Utrecht: LOT.

Gauvain, J. L.; Lamel, L. and Adda, G. (2002). The LIMSI Broadcast News transcription system. *Speech Communication*, 37(1-2), pp. 89-108.

Grepl, M.; Hladká, Z.; Jelínek, M.; Karlík, P.; Krčmová, M.; Nekula, M.; Rusínová, Z. and Šlosar, D. (1996). *Příruční mluvnice češtiny*. Prague: Nakladatelství Lidové noviny.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama, K. Maekawa (Eds.), *Proceedings of the 1st Session of the 10th International Symposium on Spontaneous Speech: Data and Analysis*. Tokyo.

Kohler K. J. (1990). Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. J. Hardcastle, A. Marchal (Eds.), *Speech Production and Speech Modelling*. Dordrecht: Kluwer Academic Publishers, pp. 69-92.

Kopřivová, M., Waclawičová, M. (2008). *Čeština v mluveném korpusu*. Prague: Nakladatelství Lidové noviny.

Krčmová, M. (2000). Termín obecná čeština a různost jeho chápání. In Z. Hladká, P. Karlík (Eds.), *Čeština – univerzália a specifika, 2*. Brno: Masaryk University, pp. 63-77.

Local, J. (2003). Phonetics and talk-in-interaction. In *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, Spain. pp. 115-118.

Ouředník, P. (2005). *Šmírbuch jazyka českého : slovník nekonvenční češtiny: 1945 – 1989*. Litomyšl: Paseka.

Pollák, P. ; Hanžl, V. ; Černocký, J. and Smrž, P. (2008). Problems and solutions in the creation of Czech and Slovak lexica for speech technology applications: General experiences and LC-Star2 lexica. In *Digital Technologies 2008*. Žilina, pp 1-5.

Psutka, J. et al. (2001). Large broadcast news and read speech corpora of spoken Czech. In *Proceedings of the 7th European Conference on Speech Communication and Technology*. Aalborg, Denmark, pp. 2067-2070.

Siemund, R. ; Höge, H.; Kunzmann, S. and Marasek, K. (2000). SPEECON - Speech data for consumer devices. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece, pp. 883-886.

Torreira, F. ; Adda-Decker, M. and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52, 201-221.

Torreira, F., Ernestus, M. (2012). Weakening of intervocalic /s/ in the Nijmegen Corpus of casual Spanish. *Phonetica*, 69, 124-148.

Volín, J. et al. (2008). Reliabilita a validita popisných kategorií v Pražském fonetickém korpusu. In M. Kopřivová, M. Waclawičová (Eds.), *Čeština v mluveném korpusu*. Prague: Nakladatelství Lidové noviny, pp. 249-254.