Univerzita Karlova
Filozofická fakulta
Ústav Českého národního korpusu

# Investigating prosody in spoken Czech: A corpus-linguistic approach

## Autoreferát disertační práce

Mgr. David Lukeš

(K prozodii mluvené češtiny metodami korpusové lingvistiky)

Disertační práce
Vedoucí práce: Mgr. Pavel Vondřička, PhD
Rok podání práce: 2022

# 1 Introduction

Czech speech recorded in naturalistic settings is available in transcribed corpora of non-trivial size (in the millions of tokens), providing a treasure trove of data which is worth exploring and experimenting with. In particular, seeing as preparing this type of corpus is very labor intensive and expensive, it is worth investigating how the data can be further enriched and analyzed in more detail using state-of-the-art speech processing tools, at a fraction of the cost of manual annotation.

In the edited volume ***Prosody in interaction*** (Barth-Weingarten, Reber & Selting 2010), Arnulf Depperman, co-author of the GAT-2 speech transcription guidelines (Selting et al. 2009), has a paper titled ***Future prospects of research on prosody: The need for publicly available corpora*** (Deppermann 2010). In it, he (unsurprisingly, given the title) advocates the need for large corpora to be compiled and made available to the academic public. I couldn't agree more with this sentiment, but as noted large and richly (e.g. prosodically) annotated are two requirements that are typically in contradiction. The present work summarizes my attempt at overcoming it.

A terminological note: the term prosody has various definitions. In the context of this study, it should be taken as encompassing all suprasegmental features of speech. The two most commonly studied subcategories under that umbrella definition are speech phenomena that are pitch-related, i.e. intonation, and phenomena which are duration-related. While the results presented in this study focus primarily on the intonation side, the underlying data yielded by the processing pipeline provides rich information that can be used for duration-focused analyses as well, laying the groundwork for studies of rhythm, speech or articulation rate, etc. Hence my frequent usage of the more general terms "prosody" or "prosodic", especially when discussing annotation.

# 2 Czech prosody in a corpus linguistics context

## 2.1 Bird's eye view of Czech intonation

Czech prosody is perhaps best-known abroad (if at all) through Czech-accented English, whose melody "typically sounds flat and monotonous to both native and proficient non-

native ears, as if signalling boredom, disinterest or lack of involvement" (Volín, Poesová & Weingartová 2015: 109). While a narrower pitch range has been identified as one of the recurring issues with L2 intonation (Mennen 2008: 55), Volín et al. go on to show that F0 in native Czech typically displays lower central tendencies than in native English (e.g. a median of 162 Hz in women and 105 Hz in men, vs. 186 Hz and 118 Hz, cf. their Table 1 on p. 112), as well as narrower ranges: an 80-percentile range of 5.2 ST in women and 6.1 ST in men for Czech, vs. 7.1 ST and 8.1 ST for English (cf. their Figure 4 on p. 114). And while it turns out that the pitch range of Czech-accented English is even narrower, leading the authors to "hypothesize that perhaps the uncertainty or even moderate anxiety associated with speaking a foreign language could enhance the tendency of Czech speakers to use narrower pitch ranges" (Volín, Poesová & Weingartová 2015: 121), it is quite clear such a tendency exists to begin with. The authors even explicitly express confidence that the results have wider implications and applications: "We believe that data provided by 32 professional speakers may serve as reference values beyond our current study." (Volín, Poesová & Weingartová 2015: 109).

So it looks like there is an objective basis to the subjective impression that Czech intonation sounds rather dull, at least compared to a language like English. This shouldn't be too surprising, evidence has been accumulating that speech communities can differ in the pitch profiles they favor, purely as a cultural phenomenon, without any underlying physiological differences (Dolson 1994), so we should have a relatively favorable prior on such a possibility. Now as for the direction in which we should expect to observe a difference, some typological differences between Czech and English can be adduced as evidence in favor of expecting more pitch variability in English.

For one thing, Czech has fixed stress, whereas English has lexical stress, which means that correctly identifying which syllable is stressed is much more important in English because it can be a differentiating factor between words. It follows from this that stressed syllables are cued via acoustic prominence, including (but not limited to) pitch manipulation. By contrast, while native speakers of Czech can typically agree with each other on which syllables are stressed, these syllables are usually not marked by any kind of acoustic prominence, unless making a deliberate emphasis, chanting etc. Acoustic measurements show they're not longer, nor louder, nor higher than neighboring unstressed syllables (Skarnitzl 2018: 213).

Perhaps more importantly though, English often relies on intonation to specify topic–focus articulation, as evidenced by the conventionalized use of italics to signal this type of emphasis in written language:

1. Alice gave the apple to ***Bob***. (The focus is on who received the apple – Bob.)
2. Alice gave the ***apple*** to Bob. (The focus is on what was given – an apple.)
3. ***Alice*** gave the apple to Bob. (The focus is on who ***gave*** the apple – Alice.)

Czech can do this too, but it has another trick up its sleeve: relatively free word order, where topic–focus can be manipulated by moving whatever should be under focus towards the end of the sentence. Which means the italics in the three English sentences above can be idiomatically "translated" just by re-arranging the words:

1. Alice dala jablko Bobovi.
2. Alice dala Bobovi jablko.
3. Jablko dala Bobovi Alice.

In summary, it seems intuitively plausible that variability of intonation should bear much less functional load in Czech than in English. In other words, pitch variation in corpora of spoken Czech is like the proverbial needle in a haystack. So how do we find some?

## 2.2 Prosogram

Figure 1 shows what conversational Czech can easily look like in terms of intonation. This type of visualization is called a prosogram and it contains a transcript of speech time-aligned on the level of ***words*** and ***phones*** with various acoustic phenomena. Of note is the faint blue curve, denoting F0 as identified via auto-correlation, and the thick black lines overlaid on top of it, which represent a stylized version of it restricted to syllabic nuclei. The stylization is an attempt to smooth over variation that is too fine-grained to be perceptually relevant, and have the visual representation more closely match the auditory impression a listener might form based on hearing this stretch of speech. Clearly, there is not much going on in this sample, intonationally speaking. Even though it's rather long, around 10 s, and punctuated by several pauses (indicated by an underscore, _, on the ***phones*** and ***words*** tiers), the tonal targets remain very level somewhere between 150 and 200 Hz (this is a female speaker). It is easy to see how such speech can be perceived as having a monotonous,

droning quality to it.

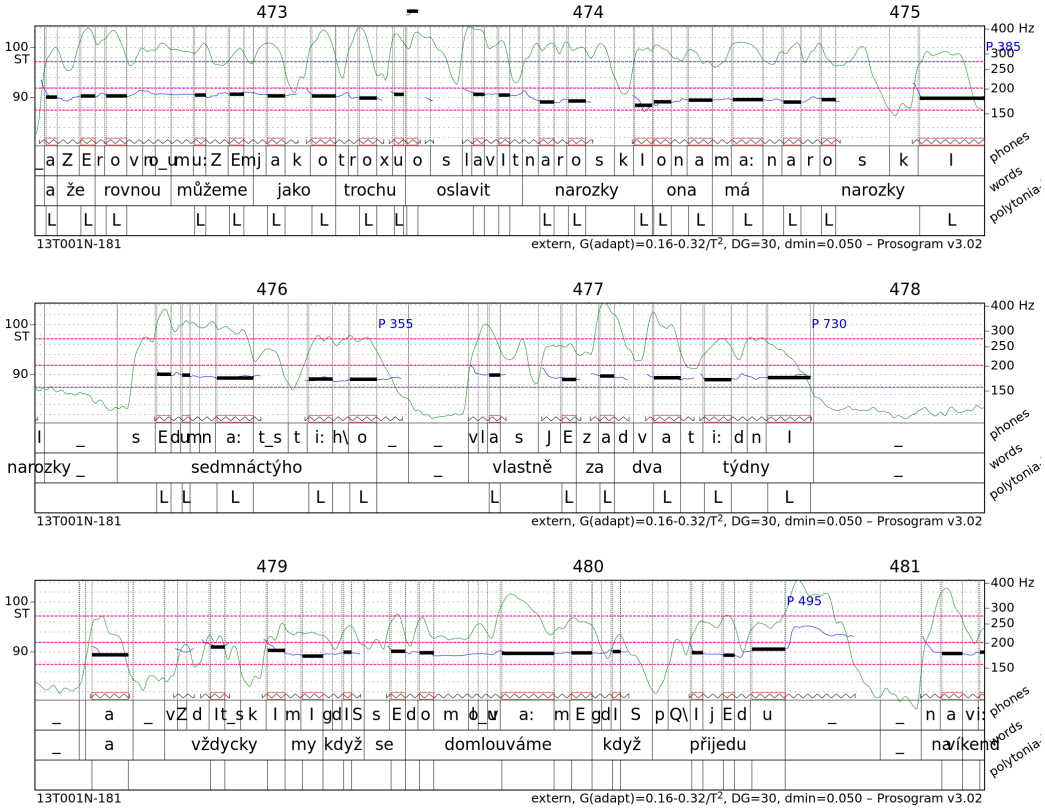


Figure 1: Prosogram of a sample of intonationally monotonous conversational Czech by a female speaker (ORTOFON v2 corpus).

However, it would be a mistake to conclude that all Czech intonation looks like this. There are occasions where speakers reach for more adventurous tonal patterns, and a lot happens in a short period of time. One such example is shown in Figure 2, which is by another female speaker, this time taken from a lecture. As indicated by some of the stylized thick black lines being inclined, the rate of F0 change in some of the nuclei is so fast that it is likely that listeners won't perceive it as a single steady tone level, but rather as a glissando. This is also reflected in the symbolic tonal transcription on the lowest *polytonia* tier: whereas in Figure 1, this tier contained a never-ending sequence of L tones (for 'low'), in Figure 2, the content is much more varied, covering rises (R) and falls (F), and reaching for the very

top (T) of the speaker's intonational range. The range itself, estimated in both cases based on the entire recording and indicated by the horizontal pink dashed lines, is also expanded compared to the first sample.

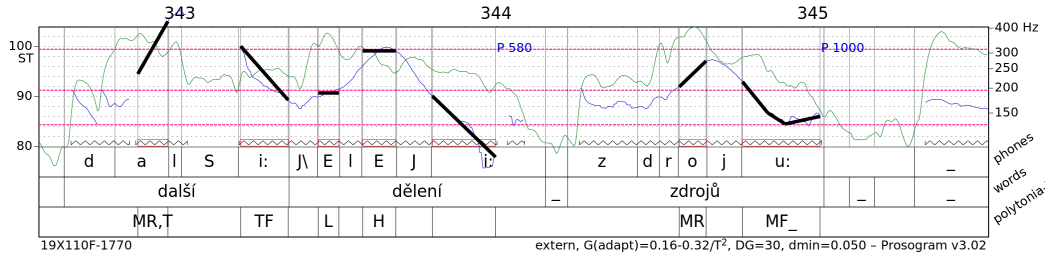


Figure 2: Prosogram of a sample of intonationally varied Czech from a lecture by a female speaker (ORATOR v2 corpus).

In other words, as with many linguistic phenomena, intonation in Czech is multi-faceted. Its internationl claim to fame, if it has any, may be monotony, and we've definitely seen a striking example of that, but we've also seen that not all of it is like that and occasionally, it can be even very lively. Wouldn't it be nice if we didn't have to sift through spoken corpora manually to find examples from either group? If we could instead slice and dice through the corpora by querying for prosodic properties, in a similar way like we use lemmatization and morphological tagging to zero in on lexical items, morphological categories or even syntactic patterns that happen to be relevant for the research project at hand? Well, this is in fact how both of the examples shown were retrieved from corpora containing millions of running words in total.

More specifically, both of the previous figures were generated using a tool called Prosogram (Mertens 2004; Mertens 2022), implemented by Piet Mertens for the Praat speech analysis environment (Boersma & van Heuven 2001; Boersma & Weenink 2022). Apart from the nice visualizations we've seen, Prosogram also reports the underlying results in a tabular format which we can inspect at leisure and map back onto the input corpus. The symbolic tonal transcript is generated via an algorithm called Polytonia (Mertens 2014), which is integrated into Prosogram and also produces output in a format suitable for further processing and analysis. This means we can cross-reference the input corpus

data with the prosodic analyses provided by Prosogram, much in the same way as we can associate the output of a part-of-speech tagger with the input text, and with all the benefits this entails.

## 2.3 Annotating the CNC spoken corpora

An additional perspective here is that I ultimately want to add prosodic annotation to the publicly released versions of the Czech National Corpus (CNC) spoken corpora available via KonText, so that other corpus users and researchers can benefit from it. What sorts of considerations apply here? What types of research should such annotation allow – nay, encourage? Which ones can it, or should it, downplay?

Let me play the devil's advocate here for starters. Why worry about annotating prosody in the context of large-scale speech corpora? Doesn't transcription already provide what most researchers need to analyze the data? And if anyone ***really*** needs to dive into the details, then access to the corresponding recordings should be luxury enough.

Or from a different angle: if we grant that some kind of prosodic annotation is worth our while, then why try to devise a theory-light, bottom-up system leveraging automatic software tools? Why not instead focus our efforts on one of the existing classification systems for prosodic phenomena, which has been battle-tested and the target audience of linguist-users is already familiar with it? In the Czech tradition, such a classical account would be František Daneš's (1957) monograph ***Intonace a věta ve spisovné češtině*** (***Intonation and the sentence in standard Czech***), which presented a taxonomy and theory of intonation patterns typically associated with different sentence and intonation unit types in Czech. Further refined over the following decades, and combined with the approach elaborated in parallel by Milan Romportl (see Petr et al. 1986: sec. 1.E.5.4 for its ultimate incarnation), the account as presented by e.g. Palková (1994) or Skarnitzl, Šturm & Volín (2016: sec. 8.5) is now broadly accepted as standard. Granted, the lexicologists among us – and there is a powerful lexicological undercurrent in corpus linguistics – would probably criticize it as being too abstract, grammar-focused, divorced from language in use, and they would be right to an extent. Luckily, a lexicon-focused take on Czech intonation exists as well: the Dictionary of Czech Phraseology and Idioms (DCPI) lists no less than 17 intonation patterns which combine in various ways with different phraseological items (Čermák

2009: sec. 2.5). Furthermore, these have already been used for annotating real-world data, specifically the Prague Spoken Corpus (Čermák, Adamovičová & Pešička 2001). So combine the two schemes somehow so that everyone is happy and be done with it?

Both Daneš's inventory and the one in DCPI have the disadvantage that the most creative linguistic work has already been done, first in establishing these inventories, deciding which contrasts to include in the classification, which to abstract over, then in applying them to the corpus material during annotation, an empirical confrontation which can engender new insights and lead to critical re-appraisals of the original theories. By contrast, once annotation is completed and the corpus is released to users, it's very tempting to reduce any kind of analysis based on them to basically accounting: which intonation pattern occurs how many times, possibly divided up into different contexts. Writing it like this is perhaps overly dismissive and unfair, so let me rephrase: such analyses definitely have their place in linguistics and can be very useful – after all, even if accounting is very mundane, it still needs to be done. But they do make it very hard to transcend any pre-established categories and discover alternative, potentially vastly better (or simply more appropriate, as language use patterns change in time) ways to structure the material at hand. This is relatively benign when the categories are mostly uncontroversial, e.g. in most cases, people would probably agree on what words a speaker said in a particular utterance, and what should therefore go into the transcript. But the effect can be far-reaching when applied to less well-traveled reaches of the language. I would argue that prosody, especially as it pertains to spontaneous language, is such a case.

These are the kinds of considerations that underlie my firm conviction that prosodic annotation in spoken corpora intended for general research should lean towards the descriptive, theory-free end of the spectrum. In other terms, it should be phonetic rather than phonological, broad rather than narrow, inclusive rather than exclusive, descriptive rather than explanatory. It should make it easier for users of the corpora to slice and dice through the data in search of meaningful patterns, whether to confirm known ones or discover new ones, without pre-imposing a possibly elegant but quite probably also restrictive system of analysis.

That being said, I also have a deeper philosophical gripe with how taxonomies (classifications, dictionaries) are typically perceived in linguistics. I think they have a tendency to lead us down the wrong path when examining how meaning works, which is a shame, because

how meaning works should be at the core of the linguistic enterprise. Let's examine that next.

# 3 How meaning works, or, the dictionary trap

## 3.1 Compositionality: The building block theory of meaning

What's wrong with how dictionaries make us perceive words, then? They make words look like building blocks. This metaphor is relatively fine for the way signifiers are put together – we string words together to make a phrase, then a sentence, then text etc. But by reifying meanings – showing us definitions alongside headwords – dictionaries fool us into thinking that signifieds *also* work like building blocks: each word "carries" a meaning, and as we snap them together to build a sentence, then in parallel with the syntactic structure thus built, a corresponding semantic structure emerges, constructed from constituent parts.

This is a **compositional** approach to meaning, traditionally associated with Gottlob Frege as the *principle of compositionality* or *Frege's principle*, although Pelletier (2001) argues it might be somewhat of a misnomer since Frege never stated such a principle in so many words; the first one to do so was Rudolf Carnap (Pelletier 2001: 89), ascribing it to Frege. Discussing what he actually calls "Frege's Principles of Interchangeability", he formulates the second one as "the sense of the whole expression is a function of the senses of the names occurring in it" (Carnap 1947: 121). A more modern, more recognizable formulation, is given e.g. by Barbara Partee in a chapter entitled *Compositionality*, which has become the dominant label for this idea (she retains however the Fregean lineage, as has become customary): "The meaning of a compound expresion is a function of the meanings of its parts and of the way they are syntactically combined" (1984: 281). As it stands, this principle underpins much of the work in semantics, both formal and less formal.

However, compositionality has also attracted criticism and controversy, with alleged counterexamples studied *ad nauseam*. One strand, the more practical-minded, down-to-earth, common sense one, is in lexicography and related fields; another is among cognitive linguists and theorists of grammar, from generative grammar to Construction Grammar (CxG), whose goals are rather loftier – to explain how language actually works in the mind.

But the long and short of it is the same in both cases: sometimes, the meaning of a whole is more than the sum of its parts. In the lexicographic case, exhibit A is typically phraseology, or more specifically idioms, and the solution is to account for multi-word units as lexical items of sorts, and build dictionaries thereof. In the cognitive/grammarian case, the poster child for this debate is so-called *logical metonymy*, in sentences like *The student begins the book*, where the intended meaning may be for instance that the student starts reading the book, but none of the *actual* words in the sentence is 'read'.

## 3.2 Meaning discrimination and information theory

How then do we explain that the meaning of the whole sometimes seems to be more than the sum of its parts? The crucial step to get this dance right is to reject the premise: *meaning does not come in parts that can be summed*. In fancy Latinate terms, meaning is not **compositional**, it's **discriminative**. For compositionality, we've been using building blocks as a metaphor. You start with *nothing*, then gradually build a structure out of blocks that fit together. Meaning is the sum of those blocks (modulo caveats above). Discrimination works the other way round: you start with *everything*, literally the whole world, and each word, or more generally, each cue, gradually refines your idea of what the speaker might be going on about, discarding hypotheses that prove untenable. Meaning is whatever is left at the end. Sometimes nothing, in which case you have to start over and figure out whether you discarded a hypothesis too eagerly (a misunderstanding), or whether it's your communication partner who's just playing fast and loose with words without trying to actually communicate something. An apt metaphor for discrimination is sculpture: you start with an amorphous lump of stone (everything in the world, not nothing), and each word is like the stroke of the mallet on the chisel, chipping away at what you can safely assume the speaker is *not* trying to say.

Lest it appear that I am claiming these insights as my own: their most vocal advocate in the present day (and certainly the person who brought them to my own attention) is Michael Ramscar (see Ramscar 2019 for a comprehensive overview and introduction, but also e.g.; Ramscar & Port 2015; Ramscar & Port 2016; Ramscar & Baayen 2013; Linke & Ramscar 2020 among others). However, the core notion that communication works discriminatively was developed by Claude Shannon and R. V. L. Hartley, whom Ramscar

calls "the founding fathers of information theory" Ramscar (2019: 10). While Shannon is indisputably the better known of the two, he studiously avoided drawing any psychological or linguistic parallels to his theory, famously stating that:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have ***meaning***; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. (Shannon & Weaver 1964: 31)

This motivated others to fill the resulting void, perhaps most notably Donald M. MacKay with his General Information Theory, which was fervently proselytized in linguistic circles by none other than Roman Jakobson (Van de Walle 2008: 114). As we'll see below, this led to some problematic misinterpretations and missteps which garnered some well-deserved criticism. Considering this, it's all the more surprising that Hartley, the other founding father (referenced right off the bat by Shannon himself) did not shy away from drawing parallels with language, and as far as I can see, he got the right idea, or at the very least he was moving in the right direction – as early as 1928:

> By successive selections a sequence of symbols is brought to the listener's attention. At each selection there are eliminated all of the other symbols which might have been chosen. As the selections proceed more and more possible symbol sequences are eliminated, and we say that the information becomes more precise. For example, in the sentence, "Apples are red," the first word eliminates other kinds of fruit and all other objects in general. The second directs attention to some property or condition of apples, and the third eliminates other possible colors. (Hartley 1928: 536)

Communication proceeds via a process of ***elimination***, or in other words, discrimination. Each word, rather than contributing an atomic block of meaning to a gradually accreting compositional structure, instead chips away at the initial amorphous lump of possibilities.

As for Jakobson's incorporation of information theory into linguistics, the result has been met with various criticisms, the most important of which is aptly summarized by

James McElvenny in an episode of his podcast *History and Philosophy of the Language Sciences*:

> But information theory was perhaps a strange savior for meaning in language. Information theory assumes that there's always a single definite message that is transmitted from sender to receiver using a fixed code. The sender and the receiver, and the context in which the message is exchanged, may have an influence on how the message is encoded and decoded, and noise may interfere with the message, but there is always a single message that is in principle recoverable. (McElvenny cca 25:00–28:00)

Ultimately, McElvenny's objection rests upon the invalid assumption that *information = meaning*, and therefore *single definite message = single definite meaning*. With a discriminative approach to meaning, we get rid of that assumption: since discrimination is something that happens in a specific person's mind, given his or her own very specific perspective and body of prior experience, including the knowledge of the language or code used to communicate the message, it follows trivially that a single definite message can discriminate various different meanings depending on who you ask.

Disagreements during the communication process can occur at two different levels: when decoding the message, and when interpreting it, i.e. using it to discriminate meaning. The first type is fully resolvable; the second one is not. In other words, even in optimal communication circumstances, where decoding error/disagreement is extremely unlikely, there is still ample room for ambiguity in meaning.

## 3.3 Practical consequences for real-life communication

Summarizing and expanding upon the most important points in the previous section: a key skill for the speaker or writer is to be able to anticipate the backgrounds and contexts of his or her audience, and also the possible misunderstandings that might arise. Since any such predictions, even if made by individuals with exceptional communication skills, finely attuned to whoever they're addressing, are liable to be wrong sometimes, it follows that dialogue should be our preferred form of interaction, if at all possible. Actual feedback is always preferable to prediction.

This ties into the notion that, since words aren't building blocks which "carry" meaning, meaning is not transferred, in the sense that there is no point at which we can declare the transfer of meaning complete; instead, it's discriminated, getting the intersubjective alignment ever closer in the ideal case. As a rule of thumb therefore, when broaching a complex issue, a few paragraphs are worth more than a single, exquisitely wrought sentence.

A mirror strategy should apply to reading: instead of agonizing over the exact meaning of a short stretch of text that is stubbornly resisting you, try to read ahead, skim, possibly even consult different sources on the same subject, and then later come back to the passage that was giving you grief. You may find it makes sense now. As for listening, I suspect the recommendations are obvious by now, but let's state them briefly: ask questions, request clarifications, engage the speaker in dialogue.

## 3.4 Historical context: The dictionary trap from Aristotle to Saussure

Even though compositionality is commonly (and possibly mistakenly) credited to Frege, the more broadly conceived dictionary trap has a distinguished intellectual history, ranging (at least in Western thought) from Aristotle to Saussure and beyond. In *De Interpretatione*, Aristotle laid the groundwork for semiotics as a theory of signs which consist of what we would today call the signifier and the signified. In Aristotle's original view, the signifier was the impression made by experience upon the mind, and it was universal: the same stimulus made the same impression on anyone. This is probably the root of the confusion: on such a view, the shorthand of saying that "words have/carry meanings" is perfectly acceptable. Since there is only ever a single, truly universal meaning behind each word, it doesn't really matter whether it is located in the mind, or the word itself. This view is mostly upheld in 17th and 18th century rationalism, although Kant attempts a more sophisticated synthesis.

When subjectivity enters the fray, it's typically at the level of languages and language communities, through the discovery of linguistic relativism in the work of figures like Herder (*Treatise On the Origin of Language*) or Wilhelm von Humboldt. These acknowledge that impressions formed by stimuli can vary between individuals, but focus primarily on what this means for studying differences between entire cultures, rather than for communication between specific individuals. When through the rationalist Port-Royal grammarians of the

17th century, Antoine Arnauld and Claude Lancelot, the tradition of Aristotelian semiotics makes it to Saussure (Joseph 2012: 144; Joseph & McElvenny 2022: 46–7), he incorporates the idea of linguistic relativism and goes even further, into a sort of linguistic isolationism: he conceives of the signified as not only arbitrary, i.e. not universal, but determined purely through relations with other signifieds within the same language system. Through his more sophisticated and abstract theory, Saussure gives the sign a second lease on life, but he also cements the idea that everything a linguist needs to know about language is in its system. The mind of the speaker is bracketed away:

> Saussure resolutely left psychology to the psychologists. Not that he dismissed it, by any means; but he'd been brought up with constant admonitions to choose a particular discipline and not stray beyond it. Saussure's expertise was as a "grammarian", as he usually called himself; any view he might venture on the psychology of language would be nothing more than opinion, not expertise, and could only damage his scholarly reputation. (Joseph & McElvenny 2022: 43)

This is a convenient and understandable move for someone who was a perfectionist at heart, and no stranger to heated, uncomfortable controversies from the start of his scientific career, with the publication of his ***Mémoire*** (Joseph 2012: 242–7).

But when faced with fundamental questions about the nature of language and meaning, I would argue leaving out the minds of actual speakers is a recipe for disaster. In terms of our running discriminative metaphor: such a decision slices off the part of the lump where all the answers are right at the start.

While Enlightenment rationalists set out on the wrong path on this issue, their empiricist counterparts were inching along in the right direction. In ***An Essay Concerning Human Understanding***, John Locke expresses the belief that the human mind starts as a blank slate (***tabula rasa***) at birth, and the concomitant worry that each of us acquires language by forming associations between sensory experiences, we might each end up with different meanings in our heads, making communication impossible. While blank slate is an oversimplification, the part about how we acquire language is spot on, including the fact that technically, we really ***do*** end up with meanings in our heads that differ from one person to another. And while this indeed leads to a myriad routine miscommunications,

an overall breakdown of communication is kept at bay by reality, which acts as pressure on intersubjective alignment.

Another way to put this is that while associations between form and meaning may be arbitrary, they are also conventional. While for a (post-)modern reader, the immediate association that the words 'arbitrary' and 'conventional' used in conjunction trigger, is with Ferdinand de Saussure, the credit for coming up with this idea of arbitrariness tempered by conventionality goes to Hugh Blair and George Campbell, two philosophers within the school of Scottish common sense realism, an 18th century offshoot of empiricism. These views were then widely taught at New England colleges in the first half of the 19th century, where American linguist William Dwight Whitney picked up on them and re-amplified them, which is how they ultimately reached Saussure (Alter 2005: 72). Unlike Saussure however, Whitney acutely realized that there is no language, and therefore no linguistics, without speakers:

> Language is, in fact, an institution—the word may seem an awkward one, but we can find none better or more truly descriptive—the work of those whose wants it subserves; it is in their sole keeping and control; it has been by them adapted to their circumstances and wants, and is still everywhere undergoing at their hands such adaptation. (Whitney 1884: 48)

Whitney wasn't the only 19th century linguist who was keenly aware that abstracting away the speakers and studying language as a reified object was an oversimplification. In a typical twist of irony, Michel Bréal – the man who coined the term 'semantics' (Bréal 1897), which as a field came to be dominated by compositionality in 20th century – actually had a much more nuanced view of how meaning works, with clear discriminative overtones. It should come as no surprise then, that figures like Whitney or Bréal are nowadays much more remembered as early precursors in the lineage of pragmatics, rather than semantics, as commonly understood today (Nerlich & Clarke 1996).

## 3.5 Concluding remarks

The earliest quote I'm aware of that puts forth an approach to meaning that is recognizably discriminative, is also rather shrewd and eloquent. Its author is Dugald Steward, another figure affiliated with Scottish common sense realism:

> [T]he function of language is not so much to ***convey*** knowledge (according to the common phrase) from one mind to another, as to bring two minds into ***the same train of thinking***; and to confine them as nearly as possible, to the same track. (Stewart 1810: 211, emphasis in the original)

In that spirit: are we on the same track now, dear reader? Or at least closer than when I kicked off by saying that ***meaning does not come in parts that can be summed***? Does ***that*** particular way of putting it make more sense now, almost in a way that makes you go "Of course, that's what he meant by saying that, it should have been obvious from the start!"? Again, the point is that it precisely ***shouldn't*** have. If it makes more sense now, it's because you now have an idea in your head, your own idea of what I was trying to convey (or, strictly speaking, confine, to adopt Stewart's vocabulary), and it is relatively easy to discriminate it by just a few words. What may have initially sounded like gibberish, or at least didn't evoke such rich connotations, may now feel like a pithy and apt formula which summarizes the essence of the discriminative approach to meaning.

But don't let that deceive you: it is only pithy and apt because you now ***already know*** what I'm trying to say. If you're tempted to believe that these few words perfectly snap together in a compositional fashion to build the intended meaning, and uttering them in front of someone new to the topic should immediately confer the same level of insight you now have, try and remember how you yourself felt when you first read them.

Coming now finally full circle – so this is perhaps the most compelling reason to avoid an existing classification, or more generally, any heavily theory-laden framework, for prosodic annotation in general-purpose spoken corpora: because meaning is not built/constructed, that's the wrong metaphor, but discriminated. As it turns out, it's also a compelling reason for having prosodic annotation ***at all***. Since transfer of meaning is never complete, only ever asymptotically approaching, each shred of evidence, each cue, helps. The transcribed words of speech are not all there is. There's much more – not just video, which is the elephant in the room, but background knowledge and shared context, which the linguist-as-analyst has precious little of compared to the actual participants of any given conversation.

# 4 Data and methodology

## 4.1 Source corpora

The two corpora used in this study, ORTOFON v2 and ORATOR v2, were built as part of the CNC project. Both feature only adult speakers, i.e. 18 years of age and older. ORTOFON (Komrsková, Kopřivová, Lukeš, Poukarová & Goláňová 2017; Kopřivová, Laubeová, Lukeš, Poukarová, et al. 2020) is a corpus of casual spoken Czech, similar in spirit and methodology to the Spoken BNC (BNC Consortium 2007; Coleman et al. 2012) or Spoken BNC2014 (Love et al. 2017). It contains spontaneous conversations mostly between family members and friends, recorded in natural, private settings; in other words, the type of language sometimes termed *intimate discourse* (Clancy 2016).

Where ORTOFON attempts to map naturally occurring private dialogues, the goal of ORATOR (Kopřivová, Laubeová, Lukeš & Poukarová 2020; Kopřivová, Laubeová & Lukeš 2021) is the same, but for monologues. The main difference is that the communication situation is symmetrical in the former (all participants are peers, equally likely to be speakers or listeners, at least in theory), whereas in the latter, it's asymmetrical (one primary, designated speaker, plus an audience, which can potentially yield a few secondary speakers). This results in a different set of speech production constraints: unlike in a dialogue, speakers don't have to manage turn-taking, but on the other hand, they have to plan ahead to sustain and organize a relatively long stretch of speaking on their own. This is likely to result in systematic differences in the use of linguistic strategies and resources, including intonation. Indeed, some of these have already been identified: Czech monologues tend to have more filled pauses and complex demonstratives, as well as higher lexical richness, than dialogues (Kopřivová, Laubeová & Lukeš 2021: sec. 5).

A little less than half of the data in ORATOR has been recorded specifically for the corpus using a procedure similar to that of ORTOFON, just under different circumstances. The rest was acquired from publicly available sources. Overall, over two thirds of the data consist of lectures, as they are the easiest material of this kind to obtain. But an effort was made to collect at least small samples of a wider range of situations, including official or ceremonial speeches, or sermons. The transcription procedure was the same as for ORTOFON, except for the phonetic tier, which was left out in this case.

As for audio quality, which is a major concern when applying automatic processing steps, it varies widely between the two corpora. ORATOR has the advantage that it focuses on settings where speech is the primary activity and most of the people present are trying to pay attention to it, which typically (though not always) results in less background noise. Some of the third-party recordings were even made using speaker-specific microphones (lavalier or otherwise), which confers exceptionally good signal-to-noise ratios in the context of the data set. On the flip side though, third-party recordings are typically available in compressed audio formats, which can affect the reliability of acoustic analyses. For F0 analysis however, and at the level of accuracy we can hope to aim for given the rest of the data, this shouldn't matter too much; it's just something to keep in mind.

First-party data, which forms all of ORTOFON and almost half of ORATOR, generally exhibits the opposite tradeoff. As mentioned, the storage format is uncompressed LPCM WAV sampled at least at 16 kHz and a bit depth of 16 bits, which is amply sufficient for F0 extraction, but the microphones are only such as afforded by a small portable recording device, and their placement tends to be only as good as the situation allows. For ORATOR, this often means that the recording is made from afar; it sounds faint and can be intermittently drowned out by noise closer by. For ORTOFON, the two major problems are ambient noise and speaker overlaps. Ambient noise comes in as many guises and flavors as you can imagine everyday situations you could have a conversation in: from occasional noises like a dog barking or a door slamming, to repeated impacts by utensils such as knives or hammers, to the sustained drone of a washing machine or car engine. As for overlaps, while algorithms exist in digital signal processing to disentangle overlapping sound sources, they generally require multiple simultaneous recordings of the scene from appropriately placed microphones (one per sound source to separate Mitianoudis 2004), which is a luxury the corpora do not provide.

## 4.2 Applying Prosogram to the corpora

Prosogram has various operating modes which have different requirements on inputs. For best results, a word- and phone-level alignment of the transcript with the recording is needed, and possibly even a grouping of the phones into syllables. However, the corpora described above only feature a text-to-sound alignment at the level of multi-word segments.

How to bridge this gap?

Fortunately, Automatic Speech Recognition (ASR) tools can be used to generate a so-called ***forced alignment***, which will do its best to estimate the location of word and phone boundaries within the segment. Two prominent speech recognition toolkits that provide this are HTK (Young et al. 2009) and Kaldi (Povey et al. 2011). However, from personal experience, using them directly can be a daunting task, especially if one is worried about optimal performance. Fortunately, more user-friendly options exist. Some of these put HTK, Kaldi or similar tools behind a web interface and offer server compute power as an additional convenience, but if you have enough computing capacity on your own, it can be useful to be able to run these tools locally, especially in more custom scenarios, or if incorporating a web service would unnecessarily complicate your data processing pipeline. One such locally installable wrapper, which delegates to Kaldi under the hood and tweaks it for the forced alignment use case, is the Montreal Forced Aligner (MFA, McAuliffe et al. 2017).

MFA, in turn, can operate in one of several ways. In general, to generate a forced alignment, one needs an acoustic model and a pronunciation dictionary. The pronunciation dictionary maps graphemes to phonemes (G2P): it establishes e.g. that when ***a*** is seen in a transcript of English speech, a pronunciation of [ə] or [e̯ɪ̯] can be expected. The acoustic model then answers questions like "What does [ə] sound like?", or "What does [ə] sound like in the context of these other two phones?" – it maps the phonetic transcript to expected acoustic patterns in the speech signal.

MFA always requires a pronunciation dictionary as input. You can either provide your own, or use MFA's conveniently bundled G2P models to generate one from your transcripts, if your language is covered. However, an acoustic model is not strictly necessary at the outset. While you can use one of the pretrained acoustic models bundled with MFA, you can also use MFA to train a new acoustic model based on your input recordings, and a forced alignment will be generated as a by-product of this training process. In the case of the present study, I went with this second option because as outlined above, there is quite a lot of variability in the acoustic quality of the recordings. Phones can exhibit different acoustic qualities depending on the recording conditions, the position of the speakers relative to the microphone, etc., and I saw no guarantee that the pretrained acoustic models would be able to encompass this variability. Conversely, in training mode, MFA offers the

option to do speaker adaptation, which I took advantage of to allow the acoustic model to adapt to each speaker within a recording separately. At the same time, the overall size of the corpora guaranteed that the acoustic models would generalize reasonably well – it is hopefully obvious why choosing to train a new acoustic model on a small data set of several dozen sentences is likely to perform worse than using an existing acoustic model.

That leaves the issue of the pronunciation dictionary. In the case of ORTOFON, this is apparently trivial – it already contains a manually prepared phonetic transcript. However, there is a hidden catch: as it strives to reflect actual pronunciation, the phonetic transcript can (and does) contain pronunciation variants for what appears on the base transcript tier as one and the same word form. The differences between them can be fairly significant, with entirely syllables being sometimes elided, as in the case of ***protože*** 'because', whose canonical pronunciation is [protoʒɛ] (three syllables), but it can undergo fairly drastic formal reduction, as is typical for high frequency words (Pluymaekers, Ernestus & Baayen 2005), resulting in pronunciations such as monosyllabic [bʒɛ] or [pr̥̩ɛ].

Now, ASR toolkits are generally able to accommodate multiple pronunciations per word form, even with weighted probabilities. Kaldi is not an exception here, and MFA exposes this functionality. However, picking out the most appropriate pronunciation variant is not something they optimize for. Their goal is ultimately to convert speech into coherent text, and phone-level alignments and pronunciation dictionaries are just an intermediary in this endeavor, a means to an end. The other component that can pick up a lot of the slack that comes with varied pronunciations is the acoustic model, and in practice, this is what Kaldi seems to prefer: providing too many dictionary variants can degrade performance, Kaldi would rather have fewer of them and account for the variation in pronunciation by making the acoustic models flexible enough to squeeze every occurrence of a given word form into one of those few dictionary variants (Lukeš, Kopřivová, et al. 2018). As far as Kaldi is concerned, this is fair game: it doesn't care about specific pronunciations, it cares about getting the words right.

This is understandable – presumably, adding variants to the pronunciation dictionary is a labor-intensive and language-specific solution, whereas making the machinery around acoustic models more flexible contributes to solving pronunciation variation in a language-agnostic and automated way, since acoustic models are bootstrapped from training data. But in this case, it's also unfortunate: in ORTOFON, a lot of manual effort has ***already***

gone into determining the specific pronunciation variant for every token, so it would be a shame to throw it all away just because Kaldi isn't really optimized for picking out the most appropriate one. Luckily, there is a way around this. Instead of building a pronunciation dictionary with variants, we can build a deterministic one and thus ensure that Kaldi always picks the variant that was specified manually. The trick is to pre-process the base transcript, so that e.g. instead of `protože`, it will contain either `protože_proto͡ɜɛ` or `protože_bɜɛ`, depending on the actual pronunciation. This will distinguish different pronunciations at the word type level in the base transcript, and consequently, instead of mapping a single word form, `protože`, to a set of competing pronunciations, the pronunciation dictionary will contain one entry for `protože_proto͡ɜɛ`, another one for `protože_bɜɛ`, etc.

As for ORATOR, there is no manual phonetic transcript, so pronunciations have to be generated. I could have used MFA's G2P models for Czech, but I ended up using the Czech phonetic transcription offered by the CorPy Python library (Lukeš 2022). The approach used by CorPy is rule-based with a system of exceptions, as opposed to using a statistical G2P model like MFA. As correspondences between Czech orthography and phonetics are relatively regular (definitely more so than in English), I deemed the predictability and introspectability of a rule-based system to be an advantage.

Having secured word- and phone-level via MFA, I then applied Prosogram and Polytonia analysis to the data. All of a speaker's segments per document were analyzed together as a unit, to make estimation of global properties such as pitch range as reliable as possible. Prosogram relies on variation in intensity for some of its calculations. As can probably be expected from the foregoing discussion of sound quality in the corpora, intensity indicators are not entirely reliable in this data set: speakers located nearer or further the microphone will tend to have higher or lower intensities on average, just by virtue of the distance, and background noise can also contribute to intensity changes. I therefore configured Prosogram to ignore intensity when segmenting the signal into nuclei, and instead fully rely on MFA's vowel segment boundaries as external segmentation. I also normalized the intensity in each segment, with the goal to amplify the quieter ones, because I had observed during experimentation that Prosogram has a tendency to skip nucleis when too quiet even when using external segmentation. As any measure that increases recall, intensity normalization has a risk of lowering precision, i.e. bringing in some garbage, but it resulted in a net improvement for specific examples I'd previously identified as problematic. Globally,

the effect was to increase the number of identified nuclei by about 15% for ORTOFON and 7.5% for ORATOR. It bears emphasizing that such normalization should really be applied *to each speech segment individually*, not to the entire recording at once. Normalization happens with respect to the loudest parts of the recording, so if a recording contains a mix of loud and quiet segments, normalizing it as a whole would not make much of a difference, because all of the segments have to fit on the same scale, so their relative intensity differences will remain unchanged. By contrast, normalizing each segment separately makes it possible to make quiet segments louder, while louder ones remain more or less as they were.

The last point where I deviated from Prosogram's suggested defaults is that I disabled automatic selection of the frequency range for F0 detection. The reason is the same as for minimizing reliance on intensity measures – sound quality issues can lead the automatic algorithm to perform suboptimally. Instead, I used fixed ranges of about 33 ST: 75–500 Hz for women, and 60–400 Hz for men. These should allow for enough headroom in the vast majority of cases.

# 5 Results and discussion

## 5.1 Cleaning up Prosogram's output

Having applied Prosogram to the ORTOFON and ORATOR corpora, I ended up with a big table of syllabic nuclei and associated information, such as the nucleus's duration, its distance in time from the previous nucleus, various indicators of stylized and unstylized F0 within the nucleus (mean, median, minimum, maximum) in Hz or ST, the amplitude of glissandos (if any), intensity, and others. At the outset, there were about 3M nuclei from ORTOFON and 2.15M nuclei from ORATOR. However, given the state of sound quality in the two corpora, these shouldn't be trusted blindly, especially for global analyses of the kind I'm about to present, where you simply can't afford to take a look at each data point individually. Some cleanup was therefore in order. The quantitative impact of the individual cleanup stages I ended up with is summarized in Tables 1 and 2.

I should point out that the cleanup steps were applied in succession, as listed in the tables, and the numbers reflect this ordering. In other words, the numbers should not be taken as straightforward indicators of the overall "usefulness" of each stage, particularly for

stages further down the pipeline: some of the material they could have in theory applied to might have already been shaved off by earlier stages. From here on afterwards, I will also start referring to ORTOFON and ORATOR as the *dialogue* and *monologue* condition respectively, as these names are more descriptive.

Details of the individual steps are given in the full text. All in all, I ended up with about 40–45% of the original dialogue data, and 90–93% of the monologue data, depending on how you count. The biggest factor contributing to the much higher mortality rate in the dialogue setting was overlaps, which are essentially absent in monologues. Note that since there was more dialogue data at the outset, after cleanup, the amount of data left happened to turn out roughly similar in each condition when measured in terms of spans (around 100,000) or words (around 1M), but quite a few more nuclei in monologue (2M) than in dialogue (1.4M). This hints at higher average word length in monologues, which is consistent with their greater lexical richness, as previously discussed.

## 5.2 Sanity checks

With cleanup out of the way, let's turn our attention to a couple sanity checks. Does that data generally look like what we would expect (Czech) intonation data to look like based on prior research, or did the uneven audio quality of the recordings lead Prosogram seriously astray? Prosogram itself computes overall prosodic profiles per speaker and document, and while I did take a look at them and they broadly seem in agreement with the results I'll present below, I won't be using them directly. The reason is simple: even though convenient, they don't take advantage of additional information I have about my data, effectively ignoring the cleanup procedure described in the previous section. While I took inspiration from the prosodic profiles as to what statistics to compute and look at, I re-computed them on my own, based on the raw per-nucleus data provided by Prosogram.

First of all, there is a general expectation that gender and age affect typical F0 values due to physiological reasons. Women tend to have a higher F0 on average, but exhibit a decreasing trend over the lifespan; men are anchored lower, and the trend is relatively flat once adulthood reached, although an uptick late in life has sometimes been observed, leading towards an overall U-shape. For an example of such data acquired in acoustically appropriate conditions, corroborating the summary presented here, see e.g. Stathopoulos,

Table 1: Summary of the impact of various cleanup stages on the number of spans, words and nuclei left in the data set, in absolute numbers. The first row is the starting point, the last row gives the final amount that was left after applying all cleanup steps, and the intervening rows specify how much got removed by the given cleanup step. Columns labeled *dialogue* refer to ORTOFON, *monologue* to ORATOR, and *total* to both data sets combined. Color coding emphasizes highest values **per column**. See text for additional details on the cleanup stages (except for first and last rows).

| | spans | | | words | | | nuclei | | |
|---|---|---|---|---|---|---|---|---|---|
| cleanup step | dialogue | monologue | total | dialogue | monologue | total | dialogue | monologue | total |
| none | 256,581 | 106,648 | 363,229 | 2,123,526 | 1,246,293 | 3,369,819 | 3,001,769 | 2,153,905 | 5,155,674 |
| prosogram | -26,397 | -804 | -27,201 | -41,398 | -5,276 | -46,674 | 0 | 0 | 0 |
| best quality | -12,052 | -4,135 | -16,187 | -101,724 | -45,294 | -147,018 | -135,410 | -76,521 | -211,931 |
| no unclear words | -10,361 | -1,127 | -11,488 | -111,331 | -14,072 | -125,403 | -140,040 | -20,206 | -160,246 |
| no overlaps | -95,382 | -517 | -95,899 | -822,533 | -4,132 | -826,665 | -1,201,510 | -6,457 | -1,207,967 |
| no low intensity | -1,622 | -498 | -2,120 | -16,086 | -6,093 | -22,179 | -20,599 | -9,312 | -29,911 |
| no suspect F0 nPVI | -7,525 | -2,992 | -10,517 | -48,199 | -25,176 | -73,375 | -58,552 | -37,570 | -96,122 |
| no Skype or car rides | -4,314 | 0 | -4,314 | -41,312 | 0 | -41,312 | -51,501 | 0 | -51,501 |
| TOTALS | 98,928 | 96,575 | 195,503 | 940,943 | 1,146,250 | 2,087,193 | 1,394,157 | 2,003,839 | 3,397,996 |

Table 2: Same as Table 1, but relative numbers, giving proportions shaved off by individual cleanup steps. Also, unlike in Table 1, color coding emphasizes highest values **across the entire table** (except for first and last rows).

| | spans | | | words | | | nuclei | | |
|---|---|---|---|---|---|---|---|---|---|
| cleanup step | dialogue | monologue | total | dialogue | monologue | total | dialogue | monologue | total |
| none | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| prosogram | -0.103 | -0.008 | -0.075 | -0.019 | -0.004 | -0.014 | 0.000 | 0.000 | 0.000 |
| best quality | -0.047 | -0.039 | -0.045 | -0.048 | -0.036 | -0.044 | -0.045 | -0.036 | -0.041 |
| no unclear words | -0.040 | -0.011 | -0.032 | -0.052 | -0.011 | -0.037 | -0.047 | -0.009 | -0.031 |
| no overlaps | -0.372 | -0.005 | -0.264 | -0.387 | -0.003 | -0.245 | -0.400 | -0.003 | -0.234 |
| no low intensity | -0.006 | -0.005 | -0.006 | -0.008 | -0.005 | -0.007 | -0.007 | -0.004 | -0.006 |
| no suspect F0 nPVI | -0.029 | -0.028 | -0.029 | -0.023 | -0.020 | -0.022 | -0.020 | -0.017 | -0.019 |
| no Skype or car rides | -0.017 | 0.000 | -0.012 | -0.019 | 0.000 | -0.012 | -0.017 | 0.000 | -0.010 |
| TOTALS | 0.386 | 0.906 | 0.538 | 0.443 | 0.920 | 0.619 | 0.464 | 0.930 | 0.659 |

Huber & Sussman (2011), Figure 1.

In our data, we unfortunately don't have age information about the monologue speakers. Therefore, Figure 3 shows a breakdown by age and gender, but only for the dialogue data. Each point is the median value of the unstylized `f0_median` measure returned by Prosogram for each nucleus, aggregated by speaker within recording. Right off the bat, we can note that men and women are fairly well separated, which seems like a low bar to clear, but it's already a good sign that Prosogram hasn't gone completely off the rails and is hopefully latching onto something real. As for the expected age-related trends, there is a hint of a negative slope in the women's data, whereas men's medians are laid out flatter. The aforementioned uptick late in life may or may not be there, the data is too sparse at this end of the range to tell reliably, especially given uneven audio quality. To give a general impression, the data set is fuzzy, to be sure, and some of the outliers should raise an eyebrow or even two as clearly suspicious, but the overall shape seems at the very least plausible.

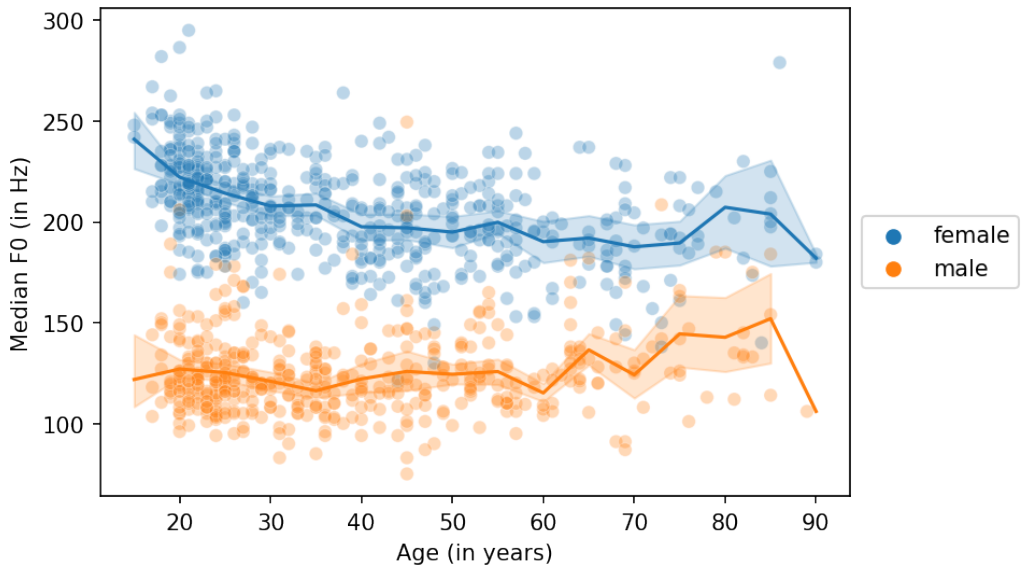


Figure 3: Median F0 per speaker in recording, in Hz, broken down by gender and age. Only covers dialogue speakers.

Not to exclude the monologue data, Figure 4 shows the distributions of median F0 before stylization per nucleus across both corpora. The upper portion is a kernel density

estimate, the lower is the empirical cumulative distribution function (ECDF). We can confirm that the distributions for men are clearly separate from those for women, with median F0 being lower for men, even in the monologue data. Additionally, we can see that the monologue distributions are somewhat shifted to the right, towards higher frequencies. This is not entirely unexpected – previous research has shown there are differences in F0 central tendency between spontaneous speech and reading e.g. in English (Hollien, Hollien & de Jong 1997), German (Jessen, Koster & Gfroerer 2005) and Czech (Skarnitzl & Vaňková 2017: 11). While our monologues are not exactly read speech, it seems plausible that they might similarly stand out. The shift observed in the previous studies under the reading condition was also rightward, towards higher F0, except in the case of German.

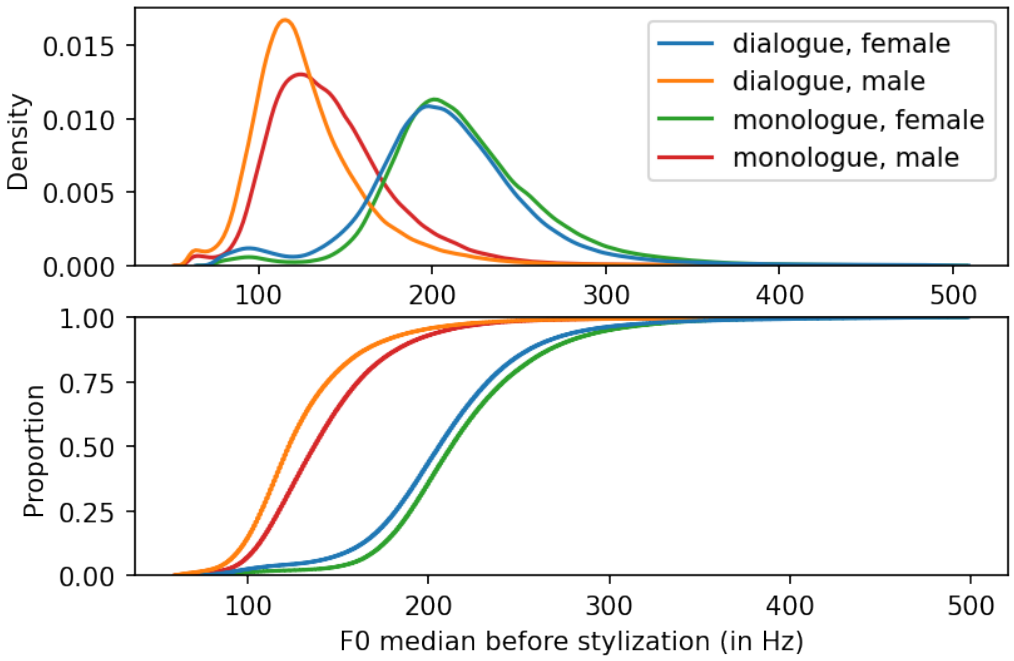


Figure 4: Distributions of median F0 per nucleus before stylization: kernel density estimate (top), empirical cumulative distribution function (bottom).

## 5.3 Glissandos

Prosogram also offers the possibility to take a look at glissandos, i.e. pitch variation ***within*** a single syllable. This is where perceptual stylization comes in particularly handy, because the ***exact*** value of F0, as extracted via auto-correlation, typically always fluctuates within a syllable, it's never a straight line. However, not all of these fluctuations are perceptually salient. Prosogram uses a glissando (and differential glissando) threshold to decide whether to stylize (model) any given stretch of F0 contour as a straight line, or as a change in pitch that can be expected to be perceptible for a typical listener. In processing the ORTOFON and ORATOR data, I stuck with the default settings for these configuration parameters, which are adaptive (Mertens 2020: 20–1). In all glissando-related analyses, I excluded syllables with F0 discontinuities, as reported by Prosogram in the `f0_discont` feature.

The distributions of glissandos, based on the `trajectory` feature reported by Prosogram, which combines the absolute values both upward and downward changes of pitch, are shown in Figure 5; they are quite similar across genders, especially in dialogue. When comparing monologues to dialogues, it appears that mild glissandos are especially symptomatic of monologues, as the place more probability mass in the left portion of the plots, close to 0. This might be explained by a presumably high incidence of continuation rises, which are typically not dramatic, but used consistently in a monologue in intonation phrase after intonation phrase, to split long utterances into more manageable chunks. At the same time though, men in monologues show a particular tendency for more pronounced glissandos: notice how the red curve in the top plot is discernibly above the other ones in the range roughly between 5 and 10 ST. Conversely, the orange curve for males in ***dialogue*** is at the opposite end, below all the other ones in this range. This discrepancy in men – a tendency for livelier intonation in monologue and duller intonation in dialogue – is something we'll come back to in the next section.

An interesting observation results from looking at the ***proportion*** of glissandos, i.e. syllable with `trajectory` greater than 0, as shown in Table 3. This proportion is higher in monologue for both genders, again possibly reflecting the regular incidence of continuation rises. Within a given setting (monologue or dialogue), the proportions are quite similar across genders, though slightly higher in both cases for men.

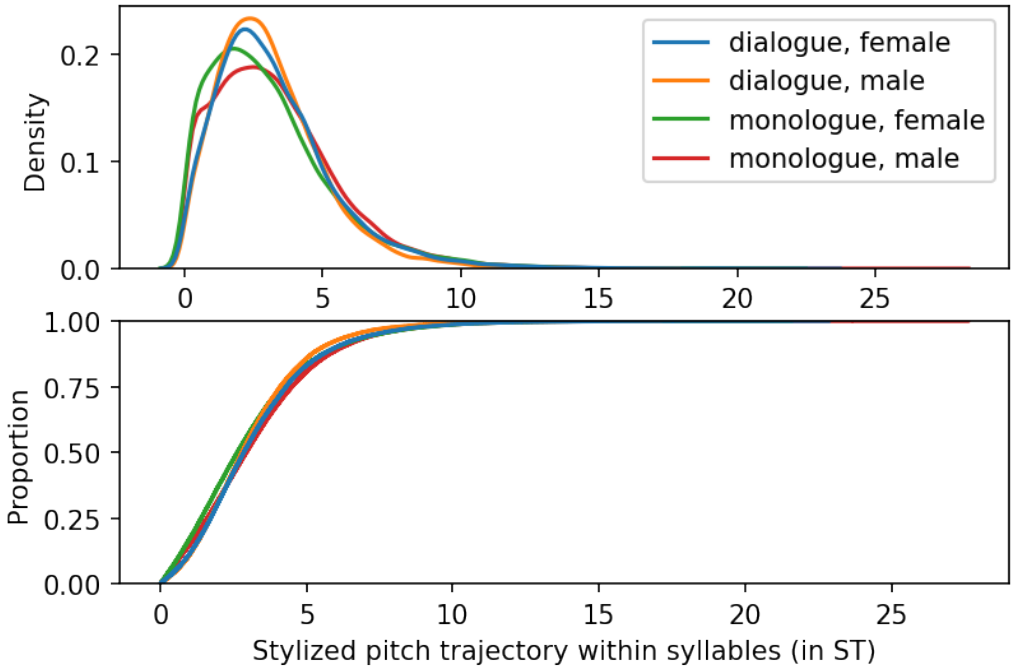


Figure 5: Distributions of glissandos (cumulative pitch trajectory per syllable) in both dialogue and monologue data: kernel density estimate (top), empirical cumulative distribution function (bottom).

Table 3: Proportion of syllables with glissandos in both dialogue and monologue data, split by gender.

| kind | gender | proportion of glissandos |
|---|---|---|
| dialogue | female | 0.0387 |
| | male | 0.0401 |
| monologue | female | 0.0454 |
| | male | 0.0488 |

## 5.4 Pitch ranges

We can now finally take a more detailed look at what factors affect F0 variation, and if Czech intonation tends to be rather monotonous by default, then which conditions – if any – counteract this tendency. A good response variable for this purpose is pitch range. A single absolute value (mean, median, or other specific quantile) lacks a point of reference: is 200 Hz a little or a lot? The answer really depends on the surrounding values, which are in turn determined to a great degree by physiological factors (cf. discussion of gender and age above). This can make it hard to tease apart what is due to conscious or unconscious decisions related to speaking style, as opposed to sheer physiology. By contrast, ranges encode variability over a span of time, irrespective of the specific absolute level at which it happened, as long as they're expressed in ST (because pitch perception is logarithmic).

An example might be helpful here: let's consider two sequences of three nuclei, with pitch targets in Hz 100–150–200 and 200–300–400, respectively. The medians are 150 and 300, but that's not exactly useful when taken out of context. Pitch ranges in Hz are 100 and 200, which makes it look like the second sequence covers more ground than the first one. However, converting to ST to account for how the ranges will actually be perceived by human ears, both ranges turn out to be identical, 12 ST. This is the level of abstraction we're looking for, one that will allow us to see past the accidents of physiology and focus on the parts of variation that speakers can and do manipulate.

What should the width of the range be then, and what unit to compute it for? In terms of width, Prosogram opts for 2nd–98th percentile pitch ranges. This is definitely an option, but narrower ranges are also used in the literature, e.g. Volín, Poesová & Weingartová (2015), whose evidence for narrower pitch ranges in Czech than in English is cited in the Introduction, uses 10th–90th percentile ranges. This seems more appropriate, given that the uneven audio quality of the recordings increases the likelihood of spurious outliers, and narrowing the range increases the chances of excluding them. As for the unit per which ranges will be computed, I opted to group nuclei into interpausal units and compute ranges for those. The minimum distance between two consecutive nuclei to be considered a pause and therefore insert an interpausal unit boundary was 350 ms, which is the pause threshold used by Prosogram (Mertens 2020: 33), and interpausal intervals of less than 6 nuclei, i.e. the lower quartile, were discarded as too short for meaningful pitch range

estimation. Another alternative would be to compute ranges per speaker in recording.

I investigated the effect of various factors on pitch range using linear models, as implemented in the ***statsmodels*** Python package (Perktold, Seabold & Taylor 2022). Where possible, I reached for a mixed effects model, specifying speaker as a random effect. In one case, the parameter estimation didn't converge, so I applied an ordinary least squares regression instead. In general, the $R^2$ of the resulting models is very low, which makes sense: much of the variation exhibited by interpausal intervals should be explained by linguistic factors, but those are completely left out at this point and left for future work. In other words, there is a lot of residual variation, and the models would work poorly when used for predicting pitch ranges. But this does not invalidate their use for analyzing the effects of those factors that ***are*** included.

Unfortunately, there is little overlap in the kinds of speaker- and document-related metadata available in the two subsets of the data defined by the dialogue and monologue conditions. The only piece of information available everywhere, and that could realistically play a role in influencing pitch range, is the speaker's gender. This is why I fitted three models: one for the entire data set, with only recording ***kind*** (dialogue vs. monologue) and ***gender*** as predictors, and one for each subcorpus defined by the ***kind*** factor, with additional factors available only in the given subcorpus.

Without further ado, Listing 1 presents the results of fitting an ordinary least squares regression model to the entire data set, with KIND and GENDER as predictors. Figure 6 then gives a visualization of the underlying distributions. The effects of both predictors, as well as their interaction, comfortably exclude 0, as can be seen in the last two columns of the table which provide a 95% confidence interval for the coefficient estimates. In other words, the contribution of the factors seems to follow a predominant direction, indicating a reliable effect. The intercept is for women under the dialogue condition, a 10th–90th percentile range of about 5.2 ST. This is virtually identical to the range reported for women by Volín, Poesová & Weingartová (2015), except in that case, the material consisted of radio news bulletins, i.e. read speech.

For easier orientation, Table 4 provides the computed predictions for the available combinations of factor levels, alongside the values from Volín, Poesová & Weingartová (2015) for reference. Please take these predictions with a grain of salt, or not literally as "predictions": as noted above, the models' $R^2$ is generally poor, so point predictions such as

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  range   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.003
No. Observations:              275358   F-statistic:                     316.4
Covariance Type:            nonrobust   Prob (F-statistic):          4.09e-205
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                         5.2111      0.015    341.629      0.000       5.181       5.241
kind[T.monologue]                -0.6169      0.024    -25.926      0.000      -0.664      -0.570
gender[T.male]                   -0.4560      0.022    -20.904      0.000      -0.499      -0.413
kind[T.monologue]:gender[T.male]  0.9367      0.031     30.623      0.000       0.877       0.997
=================================================================================================
```

Listing 1: Ordinary least squares regression model of `pitch range ~ kind + gender` in the full data set.

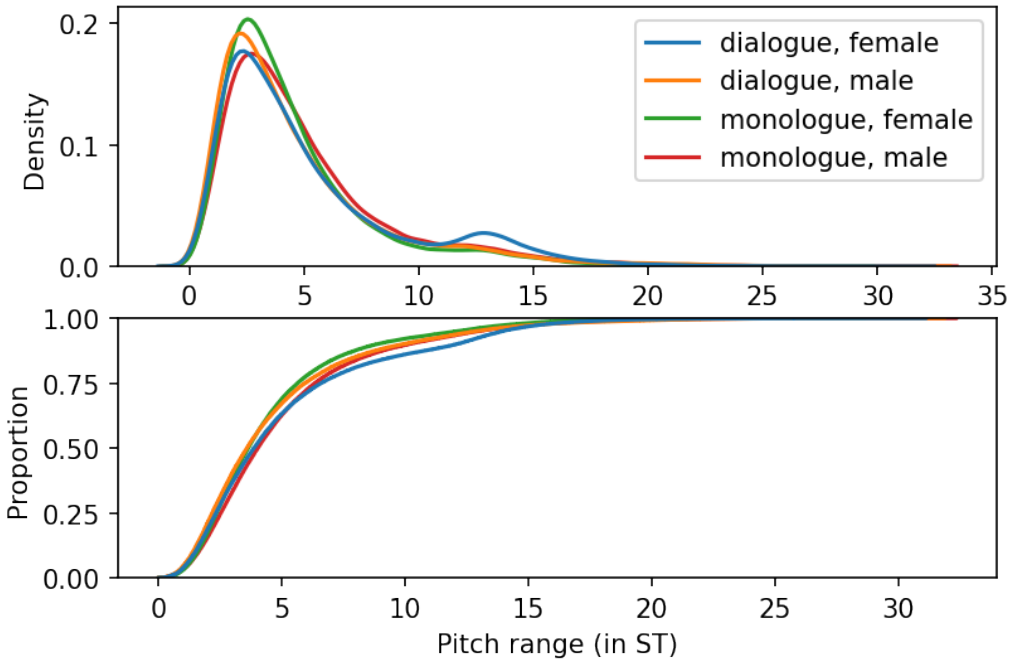


Figure 6: Distributions of interpausal pitch ranges in both dialogue and monologue data: kernel density estimate (top), empirical cumulative distribution function (bottom).

these actually hide a great amount of fuzziness. The reason I'm showing them at all is that they allow for more intuitive comparisons than the individual contrasts outputted by the model. They make it easier to see that the only condition (of those listed) where men tend towards a narrower pitch range than women is Czech dialogue. In all other conditions, including monologue from the present data set, their pitch range tends to be wider. The differences between genders amount to about 0.5 ST under the conditions investigated in the present study, and about 1 ST in the conditions investigated by Volín et al. Strikingly, the expected pitch range for women in Czech dialogue is very similar to that of men in Czech monologue, and conversely. A possible interpretation here is that women's higher pitch range in private conversations creates a gender stereotype which they are actively trying to shed in more formal settings. By contrast, men rouse themselves to use livelier intonation because they realize they don't make enough of an effort in casual speech, and aim to improve upon that baseline when addressing an audience. A similar case could be made for the differences found in Czech read speech by Volín et al., though the effect is much more pronounced, and clearly none of the Czech conditions comes even close to the ranges found in BrE read speech, male or female. However, such interpretations come with the caveat of being of course highly speculative.

Table 4: Predicted pitch range values in ST for various combinations of the KIND and GENDER factors. Where KIND is Czech monologue or dialogue, this is based on data from this study; where KIND is Czech read or BrE read, the data comes from Volín, Poesová & Weingartová (2015).

| gender ↓ kind → | Czech dialogue | Czech monologue | Czech read | BrE read |
|---|---|---|---|---|
| female | 5.21 | 4.59 | 5.2 | 7.1 |
| male | 4.76 | 5.07 | 6.1 | 8.1 |

For this summary, I'm leaving out the results of analyzing the monologue data separately, as they were less interesting. Instead, let's directly take a look at the model which focuses on the dialogue subcorpus. This is also a mixed effects model, with GENDER, CHILDHOOD REGION OF RESIDENCE and AGE as fixed effects, including an interaction between GENDER and AGE, and SPEAKER as a random effect. The CHILDHOOD REGION OF RESIDENCE is

not based on current or historic administrative subdivisions of the Czech Republic, but on the domains of occurrence of traditionally established dialects of Czech. Results of the fit are summarized in Listing 2 and again, we see a confirmation of our previous observation that men exhibit narrower pitch range in dialogue, although the exact effect sizes come out somewhat different. But the data sets differ (more specifically, the latter is only a subset of the former) and the 95% confidence intervals for the coefficients are quite wide in both cases, the results should be seen as compatible. There's also a relatively weak but apparently reliably positive correlation between pitch range and age in men: they seem to gradually increase it, ever so slightly, over the lifespan; the projected difference amounts to about 0.42 ST between a 20-year-old and 50-year-old.

As for CHILDHOOD REGION OF RESIDENCE, my prior expectations – based purely on my subjective experience as a native speaker of Czech living in the Czech Republic – were that speakers from the east of the country, i.e. Moravia and Silesia, might have somewhat wider pitch ranges. Some of this might be due to contact with Polish which, as noted in the Introduction, is not stereotypically known for dull intonation patterns, even though like Czech, it also has fixed stress. The SLEZSKÁ region in particular is under heavy influence from Polish, lacking phonemic vowel length contrasts and shifting stress to the penultimate syllable like Polish does, so it seems likely that other prosodic features would follow. And this is indeed what the data suggests: having spent one's childhood in an eastern region of the Czech Republic seems to increase the likelihood of a wider pitch range (by decreasing effect size: SLEZSKÁ, ČESKO-MORAVSKÁ, VÝCHODOMORAVSKÁ, all with 95% confidence intervals excluding 0). The one exception is STŘEDOMORAVSKÁ, which is home to the second largest city of the Czech Republic, Brno. There is also one apparent exception in the other direction: the SEVEROVÝCHODOČESKÁ region, which is technically part of Bohemia, but the explanation here might be that this is another region with close ties to Poland across the border. This is actually the region with the largest and most reliable effect size.

As mentioned at the outset, the foregoing analyses completely leave out any linguistic factors for the time being – from phonetic to lexical to syntactic to text- or discourse-related, semantic or pragmatic. Some of these may be straightforward, e.g. various types of questions make use of intonation in different ways, but probably always leading to an extended pitch range to accommodate the pattern. Others may be harder to operationalize or even pinpoint. But they are definitely worth exploring, as are more fine-grained situational or

```
                    Mixed Linear Model Regression Results
========================================================================================
Model:                      MixedLM       Dependent Variable:       range
No. Observations:           119693        Method:                   REML
No. Groups:                 926           Scale:                    14.4776
Min. group size:            1             Log-Likelihood:           -330769.0909
Max. group size:            1147          Converged:                Yes
Mean group size:            129.3
----------------------------------------------------------------------------------------
                                    Coef.  Std.Err.   z    P>|z| [0.025 0.975]
----------------------------------------------------------------------------------------
Intercept                            4.841    0.169 28.717 0.000  4.511  5.172
gender[T.male]                      -0.902    0.195 -4.630 0.000 -1.285 -0.520
reg_childhood[T.pohraničí moravské]  0.242    0.181  1.337 0.181 -0.113  0.597
reg_childhood[T.pohraničí české]     0.318    0.172  1.852 0.064 -0.019  0.654
reg_childhood[T.severovýchodočeská]  0.471    0.177  2.656 0.008  0.123  0.818
reg_childhood[T.slezská]             0.405    0.167  2.421 0.015  0.077  0.733
reg_childhood[T.středomoravská]      0.189    0.162  1.165 0.244 -0.129  0.507
reg_childhood[T.středočeská]         0.187    0.160  1.173 0.241 -0.126  0.501
reg_childhood[T.východomoravská]     0.354    0.171  2.069 0.039  0.019  0.690
reg_childhood[T.západočeská]         0.207    0.169  1.221 0.222 -0.125  0.538
reg_childhood[T.česko-moravská]      0.390    0.173  2.251 0.024  0.050  0.729
age                                  0.002    0.003  0.533 0.594 -0.004  0.008
gender[T.male]:age                   0.014    0.005  3.015 0.003  0.005  0.023
Group Var                            1.177    0.017
========================================================================================
```

Listing 2: Mixed effects model of `pitch range ~ gender * age + reg_childhood` in the dialogue subset of the data.

paralinguistic annotations available in the source corpora. These don't label the recording as a whole, they're instead linked to specific time intervals and provide information e.g. about laughter (standalone or combined with speech), emphasis or background noise. I would be surprised if proximity or overlap with at least some of these did not affect intonation to some degree.

## 5.5 Discussion and future work

The foregoing analyses are obviously just the tip of the iceberg – so much more can be done with this data, either extending and refining the angles that have been presented above, but also taking the analyses in entirely new directions. Future work should definitely include a proper comparison with English. In the present study, only a fleeting comparison was made via data from read BrE courtesy of Volín, Poesová & Weingartová (2015). Yet, the audio edition of Spoken BNC (Coleman et al. 2012) is available for download and could be processed in much the same way as the two Czech corpora used in this study.

I have actually been trying to look into this, but applying a comparable pipeline to the Audio BNC has proved troublesome so far. The data is relatively hard to work with: for one thing, the recordings are from the late 80s and early 90s, so the audio quality is only as good as portable recording devices allowed back then. But more importantly, the alignments were done *post hoc*, some 20 years later, based on separate archives containing the recordings on the one hand (the tapes had been in custody of the British Library Sound Archive), and the transcripts as published in the BNC on the other. This means that unlike in the case of ORTOFON and ORATOR, where a manual and verified span-level alignment is available, the alignment process for the Audio BNC starts with full transcripts and recordings. Inevitably, mismatches happen: the wrong recording gets paired with the wrong transcript because of faulty metadata, the transcript has parts missing that are actually present in the recording, or even the other way round. The forced aligner then does come up with an alignment (it always does – that's why it's called a *forced* aligner), but it's rubbish, nothing you can rely on for subsequent analyses. To match this with appropriate speaker metadata, contained e.g. in the XML edition of the BNC, is another sizable challenge, prone to error. I'm gradually attempting to sort or at least mitigate all of these issues, but I'm wary of trusting the results too blindly. Unlike ORTOFON and

ORATOR, this is data I barely know, so it's harder to spot systemic issues.

And of course, by this point, the original BNC data is quite old. If the audio recordings for the Spoken BNC2014 (Love et al. 2017) ever become available, it would definitely make sense to use ***those*** for comparison, instead of, or in addition to, the original Audio BNC. Hopefully they might be more reliable and easier to work with.

But beyond that, pitch range is just one of the possible ways to operationalize what we mean by intonational variability. It would be useful to fill in the current picture with additional perspectives, finding out ways that different speaking styles leverage various aspects of intonation differently. Some possibilities have been sketched at the end of the previous section, in terms of exploring more fine-grained linguistic and paralinguistic factors, instead of just speaker- and document-level metadata. Another avenue that has been explored by some empirical studies of intonation is the clustering of pitch patterns (Pęzik 2018; Raškinis & Kazlauskienė 2013; Volín & Bořil 2014). But prosody is more than just intonation – another suprasegmental feature that would be interesting to examine is word or phone durations, speech rate, or timing in general. Such an analysis might even be somewhat more reliable as it only relies on the forced alignment generated by MFA, not the F0 data provided by Prosogram.

Given the current state of the data, it is relatively cumbersome to correlate the prosodic annotation with other information also available in the corpora, be it simply n-gram context, or other annotation layers, like morphological annotation. Yet combined, easy access to all of these facets of information would open up a host of new possibilities. This is an area I would like to focus on in the near future. For one thing, this would allow building linear models which include linguistic predictors, as opposed to just metadata-based predictors. This could increase the proportion of variation explained by the models, or in other words, improve our understanding of which factors influence pitch range, or any other prosody-related response variable one might care to select.

But even more importantly, it's an important stepping stone for ultimately making prosodic annotation available to all CNC users, in the public versions of the spoken corpora accessible via KonText. This comes with its own challenges: KonText uses the Manatee corpus search engine (Rychlý 2007) as its backend, and Manatee's corpus storage and indexing format is word-based. More specifically, it requires a single tokenization, and that tokenization is intended to be roughly word-level. By contrast, MFA + Prosogram provide

us with information at multiple levels which go below that of the word: syllables and even individual phones. While it would be in theory possible to use a syllable- or phone-level tokenization in Manatee, it's not really practical: the search and indexing algorithms are not really designed for such a minute tokenization, corpora would quickly grow large and unwieldy (what counts is the number of tokens, and in this case, each phone would be a separate token), and most importantly, searching anywhere above the phone-level in a corpus prepared in this way would be extremely cumbersome in terms of query syntax.

In terms of prior art in the Czech context, I like the approach used by the Olomouc Spoken Corpus, as presented e.g. in (Pořízka 2009). This corpus uses arrows in the transcription to indicate pitch movements in a descriptive fashion, as well as three levels of pause symbols and emphasis markers. All of these symbols are added manually, which adds to the burden of transcribers and may negatively impact reliability, but overall, when manually annotating intonation, I find this descriptive, theory-agnostic approach approach preferable to trying to stick to the Daneš/Romportl/Palková analytic framework and classification discussed in Section 2.3, which is what another Czech corpus with prosodic annotation, the DIALOG corpus of TV debates, does (Čmejrková, Jílková & Kaderka 2004: sec. 2.2.2). On the other hand, a very nice feature of the DIALOG corpus is that it also provides visualizations of the intonation contours generated on-the-fly, in a strategy similar to Pęzik. While very useful for digging deeper into the elements of an already retrieved concordance, this information can't however be used for searching the corpus, as noted above.

# 6 Bibliography


Alter, Stephen G. 2005. *William Dwight Whitney and the Science of Language*. Baltimore: Johns Hopkins University Press. https://doi.org/10.1353/book.60328.

BNC Consortium. 2007. The British National Corpus, XML Edition. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2554.

Barth-Weingarten, Dagmar, Elisabeth Reber & Margret Selting (eds.). 2010. *Prosody in interaction* (Studies in Discourse and Grammar 23). Amsterdam ; Philadelphia: John Benjamins Pub. Co.

Bigi, Brigitte. 2015. SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. 108.

Boersma, Paul & Vincent van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot international* 5(9/10). 341–347.

Boersma, Paul & David Weenink. 2022. Praat: Doing phonetics by computer. https://www.praat.org.

Bréal, Michel. 1897. *Essai de Sémantique (Science des significations)*. Hachette.

Carnap, Rudolf. 1947. *Meaning and necessity: A study in semantics and modal logic*. Chicago, IL: University of Chicago Press.

Chersoni, Emmanuele, Alessandro Lenci & Philippe Blache. 2017. Logical Metonymy in a Distributional Model of Sentence Comprehension. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, 168–177. Vancouver, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-1021.

Clancy, Brian. 2016. *Investigating intimate discourse: Exploring the spoken interaction of families, couples and friends* (Domains of Discourse). London ; New York: Routledge.

Coleman, John, Ladan Baghai-Ravary, John Pybus & Sergio Grau. 2012. Audio BNC: The audio edition of the Spoken British National Corpus. Oxford: Phonetics Laboratory, University of Oxford. http://www.phon.ox.ac.uk/AudioBNC.

Daneš, František. 1957. *Intonace a věta ve spisovné češtině* (Studie a Práce Linguistické). Vol. II. Praha: Nakladatelství Československé akademie věd.

Deppermann, Arnulf. 2010. Future prospects of research on prosody: The need for publicly available corpora. In Dagmar Barth-Weingarten, Elisabeth Reber & Margret Selting (eds.), *Prosody in interaction* (Studies in Discourse and Grammar 23), 41–7. Amsterdam ; Philadelphia: John Benjamins Pub. Co.

Dolson, Mark. 1994. The Pitch of Speech as a Function of Linguistic Community. *Music perception: An interdisciplinary journal*. University of California Press 11(3). 321–331. https://doi.org/10.2307/40285626.

Grabe, Esther & Ee Ling Low. 2002. Durational Variability in Speech and the Rhythm Class Hypothesis. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory Phonology 7*, 515–546. Berlin, New York: Mouton de Gruyter.

Grieve, Jack. 2021. Observation, experimentation, and replication in linguistics. *Linguistics*. https://doi.org/10.1515/ling-2021-0094.

Hart, J. T. 't, R. Collier & A. Cohen. 1990. *A Perceptual Study of Intonation: An Experimental-*

*Phonetic Approach to Speech Melody*. Cambridge: CUP.

Hartley, R. V. L. 1928. Transmission of information. *The bell system technical journal* 7(3). 535–563. https://doi.org/10.1002/j.1538-7305.1928.tb01236.x.

Hirst, Daniel & Albert di Cristo (eds.). 1998. *Intonation Systems: A Survey of Twenty Languages*. Cambridge: CUP.

Hirst, Daniel & Robert Espesser. 1993. Automatic Modeling of Fundamental Frequency Using a Quadratic Spline Function. *Travaux de l'institut de phonétique d'aix* 75–85.

Hollien, Harry, Patricia A. Hollien & Gea de Jong. 1997. Effects of three parameters on speaking fundamental frequency. *The journal of the acoustical society of america*. Acoustical Society of America 102(5). 2984–2992. https://doi.org/10.1121/1.420353.

Jessen, Michael, Olaf Koster & Stefan Gfroerer. 2005. Influence of vocal effort on average and variability of fundamental frequency. *International journal of speech, language and the law* 12(2). 174–213. https://doi.org/10.1558/sll.2005.12.2.174.

Joseph, John & James McElvenny. 2022. Ferdinand de Saussure. In James McElvenny (ed.), *Interviews in the history of linguistics*, vol. I, 41–9. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.7092391.

Joseph, John Earl. 2012. *Saussure*. 1st ed. Oxford ; New York: Oxford Unievrsity Press.

Linke, Maja & Michael Ramscar. 2020. How the Probabilistic Structure of Grammatical Context Shapes Speech. *Entropy* 22(1). 90. https://doi.org/10.3390/e22010090.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International journal of corpus linguistics*. John Benjamins 22(3). 319–344. https://doi.org/10.1075/ijcl.22.3.02lov.

Lukeš, David. 2022. CorPy 0.4.1. Praha. https://corpy.rtfd.io.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, 498–502. ISCA. https://doi.org/10.21437/Interspeech.2017-1386.

McElvenny, James. Karl Bühler's Organon model and the Prague Circle. https://hiphilangsci.net/2022/01/01/podcast-episode-21/.

Mennen, Ineke. 2008. Phonological and phonetic influences in non-native intonation. In *Phonological and phonetic influences in non-native*, 53–76. De Gruyter Mouton. https://doi.org/10.1515/9783110198751.1.53.

Mertens, Piet. 2004. The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In *Proceedings of Speech Prosody 2004*, 4. Nara, Japan.

Mertens, Piet. 2014. Polytonia: A system for the automatic transcription of tonal aspects in speech corpora. *Journal of speech sciences* 4(2, 2). 17–57. https://doi.org/10.20396/joss.v4i2.15053.

Mertens, Piet. 2020. *Prosogram user's guide*. https://sites.google.com/site/prosogram/home.

Mertens, Piet. 2022. Prosogram + Polytonia. https://sites.google.com/site/prosogram/.

Mitianoudis, Nikolaos. 2004. *Audio Source Separation Using Independent Component Analysis*. Queen Mary, University of London PhD thesis.

Moore, Roger K. 2005. Results from a survey of attendees at ASRU 1997 and 2003. In *Interspeech 2005*, 117–120. ISCA. https://doi.org/10.21437/Interspeech.2005-82.

Müllerová, Olga. 2022. *Dialog a mluvená čeština: Výbor z textů* (Sociolingvistická edice). (Ed.) Jana Hoffmannová, Lucie Jílková & Petr Kaderka. Praha: Nakladatelství Lidové noviny.

Nerlich, Brigitte & David D. Clarke. 1996. *Language, action, and context: The early history of pragmatics in Europe and America, 1780-1930* (Amsterdam Studies in the Theory and History of Linguistic Science volume 80). Amsterdam Philadelphia: John Benjamins publishing company.

Palková, Zdena. 1994. *Fonetika a fonologie češtiny*. Praha: Karolinum.

Partee, Barbara H. 1984. Compositionality. In F. Landman & F. Veltman (eds.), *Varieties of formal semantics* (GRASS 3), 281–311. Dordrecht: Foris.

Pelletier, Francis Jeffry. 2001. Did Frege Believe Frege's Principle? *Journal of logic, language and information* 10(1). 87–114. https://doi.org/10.1023/A:1026594023292.

Perktold, Josef, Skipper Seabold & Jonathan Taylor. 2022. Statsmodels 0.13.2.

Petr, Jan, Miloš Dokulil, Karel Horálek, Jiřina Hůrková & Miloslava Knappová (eds.). 1986. *Mluvnice češtiny*. Vol. 1. Praha: Academia.

Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen. 2005. Lexical frequency and acoustic reduction in spoken Dutch. *The journal of the acoustical society of america* 118(4). 2561–2569. https://doi.org/10.1121/1.2011150.

Port, Robert F. & Adam P. Leary. 2005. Against formal phonology. *Language* 81(4). 927–964. https://doi.org/10.1353/lan.2005.0195.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, et al. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 4. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society.

Pořízka, Petr. 2009. Olomouc Corpus of Spoken Czech: Characterization and main features of the project. *Linguistik online* 38(2).

Pęzik, Piotr. 2018. Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix. In *Proceedings of LREC 2018*, 4297–4300. Miyazaki, Japan: ELRA.

Ramscar, Michael. 2019. Source codes in human communication. ms. http://arxiv.org/abs/1904.03991. (17 May, 2020).

Ramscar, Michael & Harald Baayen. 2013. Production, comprehension, and synthesis: A communicative perspective on language. *Frontiers in psychology* 4. https://doi.org/10.3389/fpsyg.2013.00233.

Ramscar, Michael & Robert Port. 2015. 4. Categorization (without categories). In *Handbook of Cognitive Linguistics*. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110292022-005.

Ramscar, Michael & Robert F. Port. 2016. How spoken languages work in the absence of an inventory of discrete units. *Language sciences* 53. 58–74. https://doi.org/10.1016/j.langsci.2015.08.002.

Raškinis, Gailius & Asta Kazlauskienė. 2013. From Speech Corpus to Intonation Corpus: Clustering Phrase Pitch Contours of Lithuanian. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, 353–363.

Rychlý, Pavel. 2007. Manatee/Bonito—A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65–70. Brno: Masaryk University.

Sabien, Duncan. 2021. Ruling Out Everything Else. LessWrong. https://www.lesswrong.com/posts/57sq9qA3wurjres4K/ruling-out-everything-else. (21

September, 2022).

Selting, Margret, Peter Auer, Dagmar Barth-Weingarten, Jörg R. Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, et al. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10. 353–402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf.

Shannon, Claude E. & Warren Weaver. 1964. *The Mathematical Theory of Communication*. Urbana: The University of Illinois Press.

Skarnitzl, Radek. 2018. Fonetická realizace slovního přízvuku u delších slov v češtině. *Slovo a slovesnost* 79(3). 199–216. https://www.ceeol.com/search/article-detail?id=687647. (19 September, 2019).

Skarnitzl, Radek & Jitka Vaňková. 2017. Fundamental frequency statistics for male speakers of Common Czech. *Auc philologica* 2017(3). 7–17. https://doi.org/10.14712/24646830.2017.29.

Skarnitzl, Radek, Pavel Šturm & Jan Volín. 2016. *Zvuková báze řečové komunikace: fonetický a fonologický popis řeči*. 1st edn. Praha: Karolinum.

Starý, Zdeněk. 1993. The forbidden fruit is the most tempting or why there is no Czech sociolinguistics. In Eva Eckert (ed.), *Varieties of Czech: Studies in Czech sociolinguistics*, 79–95. Amsterdam, Atlanta, GA: Rodopi.

Stathopoulos, Elaine T., Jessica E. Huber & Joan E. Sussman. 2011. Changes in Acoustic Characteristics of the Voice Across the Life Span: Measures From Individuals 4–93 Years of Age. *Journal of speech, language, and hearing research*. American Speech-Language-Hearing Association 54(4). 1011–1021. https://doi.org/10.1044/1092-4388(2010/10-0036).

Stewart, Dugald. 1810. *Philosophical Essays*. Edinburgh: Creech.

Taleb, Nassim Nicholas. 2018. *Skin in the game: Hidden asymmetries in daily life*. First edition. New York: Random House.

Van de Walle, Jürgen. 2008. Roman Jakobson, cybernetics and information theory: A critical assessment. *Folia linguistica historica*. De Gruyter Mouton 42. 87–123. https://doi.org/10.1515/FLIH.2008.87.

Volín, Jan & Tomáš Bořil. 2014. General and Speaker-specific Properties of Fo Contours in Short Utterances. *Acta universitatis carolinae philologica* (1). 9–20.

Volín, Jan, Kristýna Poesová & Lenka Weingartová. 2015. Speech Melody Properties in English, Czech and Czech English: Reference and Interference. *Research in language* 13(1). 107–123. https://doi.org/10.1515/rela-2015-0018.

Whitney, William Dwight. 1884. *Language and the study of language*. London: Trübner.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, et al. 2009. *The HTK Book (for HTK Version 3.4)*. Microsoft Corporation/Cambridge University Engineering Department.

Yudkowsky, Eliezer. 2015. *Rationality: From AI to Zombies*. Machine Intelligence Research Institute.

Čermák, František (ed.). 2009. *Slovník české frazeologie a idiomatiky 4: Výrazy větné*. 1st edn. Praha: LEDA.

Čermák, František, Anna Adamovičová & Jiří Pešička. 2001. *PMK (Pražský mluvený korpus): Přepisy nahrávek pražské mluvy z 90. let 20. století*. Praha: Ústav Českého národního korpusu FF UK. http://www.korpus.cz.

Čmejrková, Světla & Jana Hoffmannová (eds.). 2011. *Mluvená čeština: hledání funkčního rozpětí*. Praha: Academia.

Čmejrková, Světla, Lucie Jílková & Petr Kaderka. 2004. Mluvená čeština v televizních debatách: Korpus DIALOG. *Slovo a slovesnost* 65(4). 243–269.

## 7 Teaching and research activities

(Not including conference presentations without a resulting proceedings paper.)

- 2015: participant at *Statistics for linguistics with R bootcamp*, taught by Stefan Th. Gries, Belgium, Université Catholique de Louvain
- 2016–present: course instructor for *Programování pro korpusovou lingvistiku: Python a NLTK I and II* at FF UK
- 2017: participant at *LERU Doctoral Summer School 2017 on Citizen Science*, Switzerland, Universität Zürich
- 2017: BA thesis opponent for *Frekvenční distribuce nominální flexe v češtině* by Vojtěch Janda
- 2019: 3 day workshop instructor for *The plumbing of corpus linguistics: A guided tour of the corpus-processing pipeline*, by invitation from Prof. Achim Rabus, Germany, Albert-Ludwigs-Universität Freiburg

- 2019: summer school instructor for ***V4Py: Gentle Introduction to Natural Language Processing and Corpus Linguistics*** (with Python), organized by Silvie Cinková from ÚFAL MFF UK
- 2019: programmer at ***<Hands:On> Digital Training Programme for Postgraduate Medievalists*** hackathon, jointly organized by Cambridge University Library and Queen Mary University London

# 8 Corpora


Benešová, Lucie, Zuzana Komrsková, Marie Kopřivová, Michal Křen, David Lukeš, Petra Poukarová & Martina Waclawičová. 2017. ORAL: korpus mluvené češtiny. Ústav Českého národního korpusu FF UK. https://korpus.cz.

Goláňová, Hana, Zuzana Komrsková, Marie Kopřivová, David Lukeš, Petra Poukarová & Martina Waclawičová. 2017. DIALEKT v1: nářeční korpus češtiny. Ústav Českého národního korpusu FF UK. https://korpus.cz.

Goláňová, Hana, David Lukeš & Martina Waclawičová. 2021. DIALEKT v2: nářeční korpus češtiny. Ústav Českého národního korpusu FF UK. https://korpus.cz.

Komrsková, Zuzana, Marie Kopřivová, David Lukeš, Petra Poukarová & Marie Škarpová. 2017. ORTOFON: korpus neformální mluvené češtiny s víceúrovňovým přepisem. Ústav Českého národního korpusu FF UK. https://korpus.cz.

Komrsková, Zuzana, David Lukeš, Petra Poukarová, Anna Řehořková & Adrian Jan Zasina. 2018. Koditex: korpus diverzifikovaných textů. Ústav Českého národního korpusu FF UK. https://korpus.cz.

Kopřivová, Marie, Zuzana Laubeová, David Lukeš & Petra Poukarová. 2019a. Korpus monologů: ORATOR. Ústav Českého národního korpusu FF UK. https://korpus.cz.

Kopřivová, Marie, Zuzana Laubeová, David Lukeš & Petra Poukarová. 2020. ORATOR v2: Korpus monologů. Ústav Českého národního korpusu FF UK. https://korpus.cz.

Kopřivová, Marie, Zuzana Laubeová, David Lukeš, Petra Poukarová & Marie Škarpová. 2020. ORTOFON v2: Korpus neformální mluvené češtiny s víceúrovňovým přepisem.

Ústav Českého národního korpusu FF UK. https://korpus.cz.

## 9 Book contributions

Cvrček, Václav, Zuzana Komrsková & David Lukeš. 2018. Rozsah registrové variability textů. In *Výzkum CPACT: Komputační psycholingvistická analýza českého textu*, 153–172. České Budějovice: Pedagogická fakulta Jihočeské univerzity v Č. Budějovicích.

Cvrček, Václav, Zuzana Laubeová, David Lukeš, Petra Poukarová, Anna Řehořková & Adrian Jan Zasina. 2020a. *Registry v češtině* (Studie z korpusové lingvistiky 27). Nakladatelství Lidové noviny.

Klimešová, Petra, Zuzana Komrsková, Marie Kopřivová & David Lukeš. 2017. Avenues for Research on Informal Spoken Czech Based on Available Corpora. In Piotr Pęzik & Jacek Tadeusz Waliński (eds.), *Language, Corpora and Cognition* (Łódź Studies in Language), vol. 51, 145–162. Frankfurt am Main: Peter Lang Edition.

## 10 Articles

Cvrček, Václav, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková & Adrian Jan Zasina. 2018a. Variabilita češtiny: multidimenzionální analýza. *Slovo a slovesnost* 79(4). 293–321.

Cvrček, Václav, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková & Adrian Jan Zasina. 2018b. From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. *Corpus linguistics and linguistic theory* 17(2). https://doi.org/10.1515/cllt-2018-0020.

Cvrček, Václav, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková, Adrian Jan Zasina & Vladimír Benko. 2020. Comparing web-crawled and traditional corpora. *Language resources and evaluation* 54(3). 713–745. https://doi.org/10.1007/s10579-020-09487-4.

Cvrček, Václav, Zuzana Laubeová, David Lukeš, Petra Poukarová, Anna Řehořková & Adrian Jan Zasina. 2020b. Author and register as sources of variation: A corpus-based study using elicited texts. John Benjamins. https://doi.org/10.1075/ijcl.19020.cvr.

Goláňová, Hana, Marie Kopřivová, David Lukeš & Martin Štěpán. 2015. Kartografické a geografické zpracování dat z mluvených korpusů. *Korpus – gramatika – axiologie* (11).

42–54.

Klimešová, Petra, Zuzana Komrsková, Marie Kopřivová & David Lukeš. 2015. Slovo to v mluvených korpusech ČNK, jeho prefixace a reduplikace. *Časopis pro moderní filologii (Journal for Modern Philology)* 97(1). 21–30. https://dspace.cuni.cz/handle/20.500.11956/96540.

Komrsková, Zuzana, Marie Kopřivová, David Lukeš, Petra Poukarová & Hana Goláňová. 2017. New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Journal of linguistics/jazykovedný casopis* 68(2). 219–228. https://doi.org/10.1515/jazcas-2017-0031.

Kopřivová, Marie, Zuzana Komrsková, David Lukeš & Petra Poukarová. 2017. Korpus ORAL: sestavení, lemmatizace a morfologické značkování. *Korpus - gramatika - axiologie* (15). 47–67.

Kopřivová, Marie, Zuzana Komrsková, Petra Poukarová & David Lukeš. 2019. Relevant Criteria for Selection of Spoken Data: Theory Meets Practice. *Journal of linguistics/jazykovedný casopis* 70(2). 324–335. https://doi.org/10.2478/jazcas-2019-0062.

Kopřivová, Marie, Zuzana Laubeová & David Lukeš. 2021. Designing a corpus of Czech monologues: ORATOR v2. *Jazykovedný časopis* 72(2). 520–530. https://doi.org/10.2478/jazcas-2021-0048.

Lukeš, David. 2018. Válka s žánry. *Studie z aplikované lingvistiky / Studies in Applied Linguistics* 9(2). 76–80. https://dspace.cuni.cz/handle/20.500.11956/104679. (8 March, 2019).

Lukeš, David, Dita Fejlová & Radek Skarnitzl. 2014. Variability of Czech Alveolar Plosives: A Locus Equation Perspective. *Acta Universitatis Carolinae Philologica, Phonetica Pragensia* XIII(1). 21–32.

Lukešová, Lucie & David Lukeš. 2021. Frekvence vyjmenovaných slov v současné češtině: korpusová studie. *Didaktické studie* 13(1). 125–156.

Šturm, Pavel & David Lukeš. 2017. Fonotaktická analýza obsahu slabik na okrajích českých slov v mluvené a psané řeči. *Slovo a slovesnost* 78(2). 99–118.

## II Conference proceedings

Fejlová, Dita, David Lukeš & Radek Skarnitzl. 2013. Formant Contours in Czech Vowels:

Speaker-discriminating Potential. In *INTERSPEECH-2013*, 3182–3186.

Klimešová, Petra, Zuzana Komrsková, Marie Kopřivová & David Lukeš. 2014a. Cože to je? K tvaru to v mluvených korpusech ČNK. In *Korpusová lingvistika Praha 2014 – Abstrakty*, 95–98. Praha: Ústav Českého národního korpusu.

Kopřivová, Marie, Hana Goláňová, Petra Klimešová, Zuzana Komrsková & David Lukeš. 2014. Multi-tier Transcription of Informal Spoken Czech: The ORTOFON Corpus Approach. In *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure* (Olomouc Modern Language Series, Vol. 4), 529–544. Olomouc: Univerzita Palackého.

Kopřivová, Marie, Petra Klimešová, Hana Goláňová & David Lukeš. 2014. Mapping Diatopic and Diachronic Variation in Spoken Czech: The ORTOFON and DIALEKT Corpora. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 376–382. Reykjavik, Iceland: European Language Resources Association (ELRA).

Kopřivová, Marie, Zuzana Laubeová, David Lukeš & Petra Poukarová. 2019b. Creating a sociologically balanced spoken corpus. In *Proceedings of the international conference «Corpus Linguistics – 2019»*, 40–47. Saint Petersburg: Saint Petersburg University Press.

Lukeš, David. 2015a. New Tools for Working with the ORAL Series Corpora of Spoken Czech: AchSynku and MluvKonk. In Katarína Gajdošová & Adriána Žáková (eds.), *Natural Language Processing, Corpus Linguistics, Lexicography*, 90–101. Bratislava, Slovakia: RAM-Verlag.

Lukeš, David, Petra Klimešová, Zuzana Komrsková & Marie Kopřivová. 2015. Experimental Tagging of the ORAL Series Corpora: Insights on Using a Stochastic Tagger. In Pavel Král & Václav Matoušek (eds.), *Text, Speech, and Dialogue* (Lecture Notes in Computer Science), 342–350. Springer International Publishing.

Lukeš, David, Marie Kopřivová, Zuzana Komrsková & Petra Poukarová. 2018. Pronunciation Variants and ASR of Colloquial Speech: A Case Study on Czech. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, et al. (eds.), *Proceedings of the Eleventh International Con-*

*ference on Language Resources and Evaluation (LREC 2018)*, 2704–2709. Miyazaki, Japan: European Language Resources Association (ELRA).