Univerzita Karlova
Filozofická fakulta
Ústav Českého národního korpusu

Investigating prosody in spoken Czech

A corpus-linguistic approach

David Lukeš

(K prozodii mluvené češtiny metodami korpusové lingvistiky)

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

Děkuju vedoucímu práce Pavlu Vondřičkovi za popohánění i podporu (nejen technickou, ovšem restart zaseklého kontejneru na vzdáleném serveru, uprostřed noci a o víkendu, nelze popsat jinak než jako obětavost); děkuju kolegům z ÚČNK, zejména kolegyním z mluvené sekce, s jejichž pomocí jsem pronikal do řemesla budování mluvených korpusů; a v neposlední řadě děkuju Lucii, Hedvice a Viktorce – za statečnost a lásku.

Prohlašuju, že jsem disertační práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Abstract

Prosody is a key aspect of spoken language, yet it is currently underrepresented in the spoken Czech corpora on offer at the Czech National Corpus. This is mainly because spoken corpora are very expensive and manual work intensive as it is, and adding more annotation manually is infeasible. The present dissertation thus charts a way to provide an automatic prosodic annotation for the spoken corpora of the CNC using the Prosogram framework, in combination with other tools and various custom postprocessing strategies and heuristics.

A case is also made in favor of theory-light, predominantly descriptive approaches when preparing general-purpose spoken corpus annotations for the consumption of the linguistics research community at large, in a variety of contexts and research tasks. This case is philosophically anchored in a discriminative approach to meaning, which is shown to be the correct, paradox-free alternative to the currently more dominant paradigm of compositionality.

Finally, a selection of results based on the Prosogram-generated annotation is presented. A particular focus is given to pitch range, which is characteristically restricted in Czech compared to other languages like English, but other features such as glissandos are also considered.

Keywords: Czech, speech, prosody, corpus linguistics, discriminative, meaning

Abstrakt

Prozodie je klíčovým aspektem mluveného jazyka, nicméně v korpusech mluvené češtiny, které jsou aktuálně v nabídce Českého národního korpusu, je reprezentována jen okrajově. Primární důvod je ten, že mluvené korpusy jsou už tak velmi náročné co se nákladů a manuální práce týče, takže přidávat další manuálně anotované prvky není schůdné. Předkládaná práce tak nabízí cestu, jak prozodickou anotaci doplnit do těchto korpusů automaticky, pomocí systému Prosogram v kombinaci s dalšími nástroji a vlastními postprocessingovými postupy a heuristikami.

Součástí teoretického zdůvodnění volby Prosogramu jako anotačního nástroje je i analýza toho, jak funguje v jazyce význam. Filozoficky je ukotvená v diskriminativním pojetí významu, které na rozdíl od aktuálně dominantního kompozičního pojetí neskýtá při důsledné aplikaci žádné paradoxy. Vyplývá z ní, že anotaci obecných mluvených korpusů, která cílí na užití širokou lingvistickou komunitou v různých kontextech a při různých výzkumných úkolech, je vhodné cílit deskriptivně, s minimální poplatností konkrétním teoriím.

Prezentované výsledky, získané pomocí zpracování Prosogramem, se soustředí zejména na intonační rozpětí, protože omezené intonační rozpětí je poměrně nápadným rysem češtiny ve srovnání s jinými jazyky, např. angličtinou. Věnujeme se nicméně i jiným rysům, např. glissandům.

Klíčová slova: čeština, mluvený jazyk, prozodie, korpusová lingvistika, diskriminativní, význam

[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.

(Dugald Stewart, Philosophical Essays *p. 211, 1810)*

Contents

1 Introduction

1.1 Why Czech?

You might be wondering why you should spend time reading about prosodic features of spoken Czech, of all things. The language is neither big enough to be globally relevant, nor small enough to be endangered and therefore intrinsically worth documenting. It's part of the mundane middle of European languages, which is fitting given its cultural affiliation to *Mitteleuropa*.

But linguistics is not only about the language, it's also about what you can do with it. And Czech happens to have an interesting advantage here: 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.

Such an endeavor inevitably comes with caveats. While speech recorded in naturalistic settings has a head start over lab speech in terms of ecological validity, the quality of the recordings is typically considerably worse. This is a limitation on the kinds of additional processing one can run without risking garbage output. Still,

it's worth pushing the limits of spoken corpus research and documenting those limitations. So even if you don't particularly care about Czech, which is perfectly understandable, the methodology, inasmuch as it employs tools broadly applicable to other languages too, should hopefully be of interest.

It might be helpful to state the perspective from which I'm writing: I have spent the last decade taking part in building spoken corpora of Czech at the Czech National Corpus (CNC)[1], together with my colleagues from the spoken corpora section and many external collaborators, and making them accessible via the KonText[2] query interface. We continuously strive to enhance them and make them more user-friendly, while supporting a wider range of possible research tasks. However, since spoken corpora are expensive and require a lot of manual work, there are limits to what can be achieved using manual annotation, both in terms of financial and human capacity. This is where automatic prosodic annotation comes in.

1.2 Why prosody?

"Automatic" prosodic annotation actually also involves quite a bit of human, manual work to set it up, as you'll undoubtedly realize once you're done reading this text. But the idea is that the amount of this work is constant: once its done, you can delegate it to machines to annotate unlimited amounts of data, at least in theory. In practice, this is never quite so, but at least, the amount of work required to process new data grows much slower than with 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***

[1] See `https://www.korpus.cz` for more information.
[2] See `https://korpus.cz/kontext`.

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.

1.3 Overview

I start by giving an overview of selected aspects of intonation in Czech in the broader context of the Czech language system, discussing the structural pressures (or lack thereof) which shape these aspects, particularly in contrast to English. I then discuss Prosogram as a framework for automatic prosodic annotation, as well as some other possible options, including traditional intonational analyses of Czech.

The next chapter, Chapter 3, is about meaning and how meaning works in language. Meaning is a fundamental concern in linguistics – arguably, perhaps *the*

fundamental concern. It deeply informs and shapes our approaches to linguistic data and its analysis. This is all the more relevant when building corpora not only for one's own use, but also for the use of other researchers. In this case, the chapter strives to make the case, starting from philosophical first principles, for preferring theory-light annotation approaches when building general-purpose spoken corpora.

The chapter looms fairly large, but this is for reasons that will hopefully become obvious when reading the chapter itself. Other parts of the text are fairly in-paradigm; this one ventures further out, so it requires much more words to make misunderstandings less likely. As to why this should be so, one of the goals of the chapter is to answer precisely that question.

While I consider the discussion of meaning a key issue and therefore devote a considerable amount of space to it, quite a bit more work, in terms of time spent, has actually gone into the data processing and analysis part. It just happens to be work that manifests primarily in computer code and data, not text. A whole section is devoted to data cleanup strategies, to make it as safe as possible to conduct bulk analyses and lower the risk of garbage in, garbage out. The heart of the analyses focuses on pitch range, because that's a relatively salient feature of Czech intonation that benefits from a study on naturalistic data: compared to other languages, e.g. English, Czech speakers typically use a noticeably restricted pitch range (see next chapter). But other facets are explored too, as well as the wide-ranging possibilities for future work, based on the automatically generated annotations.

As for the sources of data used, they are two corpora of spoken Czech I've taken part in building in the past decade at the CNC: ORTOFON v2 (Kopřivová, Laubeová, Lukeš, Poukarová, et al. 2020) and ORATOR v2 (Kopřivová, Laubeová, Lukeš & Poukarová 2020). Both contain primarily spontaneous language, but while the former collects private dialogues, the latter contains public or semi-public

monologues. As we will see such stylistic diversity will allow us to draw interesting comparisons and contrasts.

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[1] 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),

[1] F0, or fundamental frequency, is the acoustic correlate of intonation.

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,[2] nor louder, nor higher than neighboring unstressed syllables (Skarnitzl 2018: 213). This is not a necessary consequence of having fixed stress – for instance, Polish has fixed stress too, and informal observations suggest that it ***does*** correlate with acoustic prominence – but it is an outcome that seems much less likely, if not outright

[2] A likely reason for this is that Czech has phonemic vowel length, so length contrasts are already leveraged by other parts of the system.

precluded, in the case of lexical stress.

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,[3] 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 2.1 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***[4] with various acoustic

[3] My own informal impression is that it does increasingly so, possibly under the influence of English, but this is hard to test empirically and personal impressions are of course very vulnerable to confirmation bias.

[4] Phonetic transcription uses the SAMPA alphabet (Wells 1997).

phenomena. Of note is the faint blue curve, denoting F0 as identified via autocorrelation, 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.

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 2.2, 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 2.1, this tier contained a never-ending sequence of L tones (for 'low'), in Figure 2.2, 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.

In other words, as with many linguistic phenomena, intonation in Czech is multifaceted. Its internationl claim to fame, if it has any, may be monotony, and we've

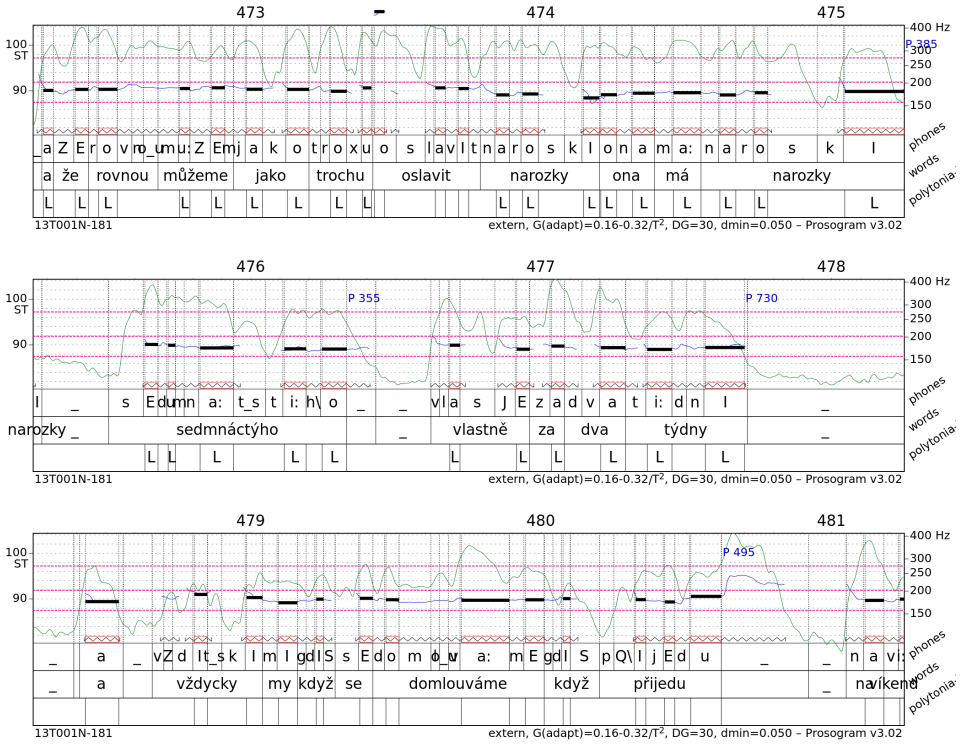

Figure 2.1: Prosogram of a sample of intonationally monotonous conversational Czech by a female speaker (ORTOFON v2 corpus).

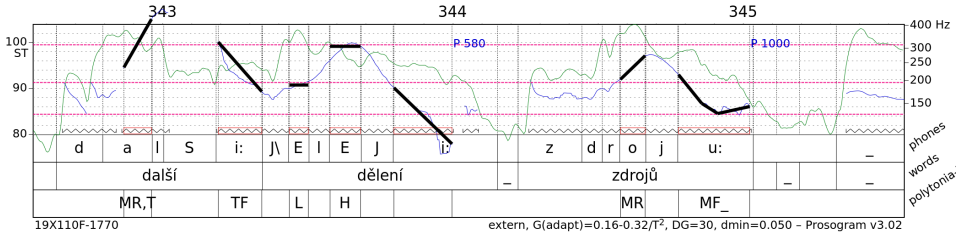

Figure 2.2: Prosogram of a sample of intonationally varied Czech from a lecture by a female speaker (ORATOR v2 corpus).

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.

I should note at this point that while Prosogram is sophisticated and polished, with extensive documentation and a detailed User's Guide,[5] it is not the only game in town for automatic analysis of F0 curves, and prior art exists which leverages alternative tools. In particular, see the pioneering work done by Piotr Pęzik on Spokes Mix[6], which provides an advanced query interface for several corpora of

[5] See Prosogram's homepage at https://sites.google.com/site/prosogram/.
[6] See http://pelcra.clarin-pl.eu/spokes2-web/.

spoken Polish and English, leveraging the Momel algorithm (Hirst & Espesser 1993) to provide a symbolic tonal transcript using the INTSINT intonation coding scheme (Hirst & di Cristo 1998). While Pęzik seems to be using a custom setup, an openly available tool that provides Momel analysis with optional subsequent coding into INTSINT codes as part of one integrated environment is Brigitte Bigi's SPPAS[7] (Bigi 2015). The common aspect of all these alternatives is that they use a two-step approach: first, the F0 curve extracted via auto-correlation is simplified (Prosogram, Momel), and that is then possibly further converted into a symbolic transcript (Polytonia, INTSINT). The first step retains more phonetic detail and enables more precise quantitative comparisons, whereas the second adds a layer of abstraction on top which can make it easier to spot generalizations or run search queries against the data. My own preference for Prosogram is motivated by several reasons. For one thing, it comes with nice visualizations out-of-the-box, which make it easy to get a quick overview and are intuitive even for non-experts (compare the thick black lines in the prosograms to the contents of the *polytonia* tier). For another, it emphasizes a *perceptual* approach to the stylization of F0 curves, which tries to distil variation that listeners are likely to notice (this tradition has its roots at the IPO in Eindhoven, see 't Hart, Collier & Cohen 1990). But the key benefit is that Prosogram takes advantage of the phone-level alignment with the transcript to only consider F0 within syllable nuclei or codas, whereas Momel makes no such distinction. Given the inconsistent audio quality of the data (see below), which can lead to F0 misdetection due to background noise, I deemed Prosogram's more restrictive approach a safer bet.

I would also like to clarify that I am writing this from a corpus linguist's perspective, rather than a phonetician's. Phoneticians routinely undertake instrumental

[7] See `http://www.sppas.org`.

analyses of intonation, whether facilitated by some of the above-mentioned tools or not, but typically on smaller-scale phonetic corpora of speech acquired in lab settings, using specific protocols and with better audio recording quality. The idea here is to start from F0 curves which are as reliable as possible, going sometimes so far as to manually correct errors due to automatic detection. By contrast, my point is that some of these analyses can also be fruitfully applied to corpora of naturalistic speech, yielding output that is admittedly messier, but at the same time, far from unmitigated garbage. Such corpora have the advantage of tending towards the unscripted end of the spontaneity continuum and should therefore be more representative of how a given language is used in the wild – in other words, the kinds of corpora that a corpus linguist investigating spoken language would intuitively reach for. Think Spoken BNC (Coleman et al. 2012) or Spoken BNC2014 (Love et al. 2017) rather than DyViS (Nolan et al. 2009). They also tend to be larger, which makes the task of wading through the data, searching for a specific pattern of interest, somewhat daunting. This is why the prospect of leveraging prosodic annotation to do some of this heavy lifting for us is enticing, especially if said annotation can be generated automatically.

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 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?

Not so fast. First of all, this type of annotation would have to be done manually, as it requires a relatively complex assessment of the utterance in context in order to select the most semantically and pragmatically adequate item on the menu, so to speak. But unfortunately, adding more layers of manual annotation quickly compounds the costs of what is already a quite expensive pipeline in the case of spoken corpora. At the CNC, spoken corpora are easily the most expensive per token of all the corpora we build and host, while also being among the smallest. Producing them is labor intensive, it takes a lot of time, money and effort from actual human beings.

Second, Daneš's heritage, while groundbreaking and apt in many ways, is also flawed in some key respects. These are mostly embodied in the choice of the word *spisovný* in the title of his monograph. I translated *spisovný* as 'standard', but a more literal translation would be 'literary', which makes the problem more obvious. Daneš's account of Czech intonation purports to describe a standard variety of Czech which is primarily couched in terms of written language. This inevitably warps the perspective and partially hamstrings the endeavor: while Daneš has many shrewd and fitting observations concerning the realities of spoken language (see Section 5.5 for a few I found particularly engaging), there is a constant nagging at the back of the reader's mind that his theory is ultimately constructed to fit much more well-behaved data than what typically occurs in spontaneous speech. Not that the observations are wrong, *per se*, they just sometimes have the slightly artificial flavor that comes with modeling overly tidy introspective data rather than the real world – as if it were an account of written-to-be-spoken rather than spoken language. As Starý (1993: 81) points out, this decision "to restrict the study of language usage to standard language" is one inherited from the Prague Linguistic Circle (PLC). Which is not to say that the PLC had a bad influence on Daneš, he made good use of

some of its other tenets, e.g. Sergei Karcevsky's notion of *compositional relationships* (see p. 159), but I think it helps explain this otherwise rather odd choice. We're all encased in a broader historical moment and tradition; at any given time, it's quite easy to take advantage of its greatest achievements, but equally difficult to spot its shortcomings and blind spots.

Third, 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.

By way of analogy, consider the subdivision into regional varieties featured in cor-

pora of spoken Czech built at the CNC. This subdivision is based on the traditional dialect regions of the Czech Republic, as established in the literature by Jaromír Bělič (Bělič 1972) and subsequently by Balhar et al. in their monumental six-volume Czech Linguistic Atlas (CLA) (Balhar et al. 1992; Balhar et al. 1997; Balhar et al. 1999; Balhar et al. 1999; Balhar et al. 2002; Balhar et al. 2005; Balhar et al. 2011). An overview of the main dialect regions, as annotated in all current CNC corpora of spoken Czech, is given in Figure 2.3. While my colleagues have done much to further refine the established dialect region boundaries, (re-)analyzing both old and new data (Goláňová & Waclawičová 2021: 503; the resulting map data has been published as Goláňová 2021), and they deserve nothing but praise for all that hard work, the fact remains that such refinememnts remain within the paradigm of an analysis which was established in a different social and historical context, primarily drawing on speakers who were born more than a century ago today. I'm not saying this analysis is no longer relevant today; it certainly is for those original speakers, who are currently featured in the DIALEKT corpus (Goláňová & Waclawičová 2019; Goláňová, Lukeš & Waclawičová 2021), and even for more contemporary recordings, I'm sure the broad strokes are still relevant and useful.

But for the linguists who use these corpora for their research, it also makes it hard to transcend the comfortable straitjacket of these existing dialect regions. For one thing, dialectology consciously strives to preserve the oldest, most conservative form of regional varieties; in English dialectology, the corresponding archetype is that of "NORM", i.e. non-mobile rural male (Chambers & Trudgill 1998: 29). But urban speech does not necessarily conform to these patterns, there is a dynamic between urban centres and rural hinterland which is left out of the picture by accepting the dialect boundaries as drawn on the map for given. For another thing, many changes happened since the end of the Second World War – increased mobility

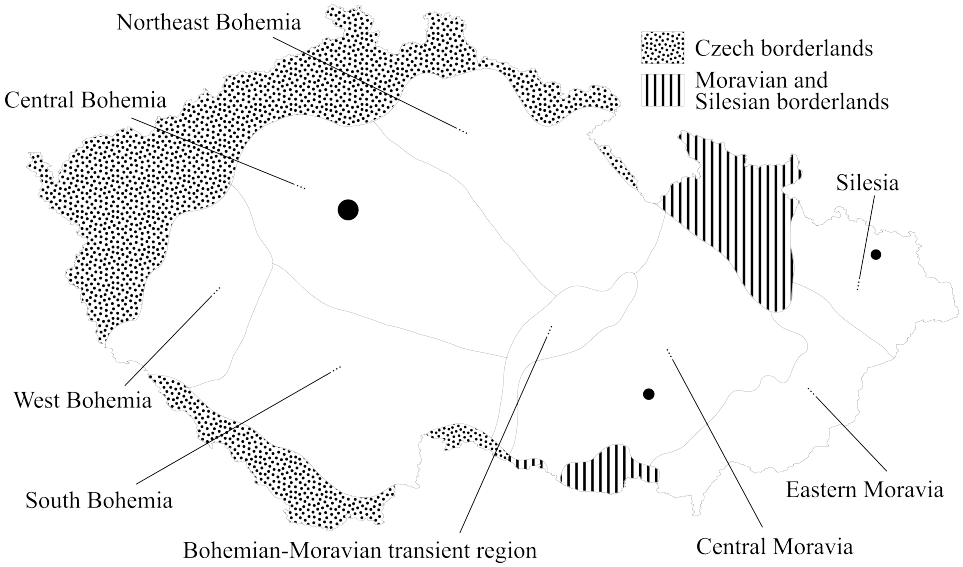

Figure 2.3: Overview of the main Czech traditional dialect regions, following Bělič and Balhar (see text). Originally prepared for Kopřivová et al. (2014).

(both geographic and social), the rise of audiovisual mass media, various political upheavals. Again, it's hard to study the impact of these developments through the prism of dialects rooted in the 19th century. Perhaps most conspicuously, the Czech, Moravian and Silesian borderlands (dotted and striped in Figure 2.3) are traditionally excluded from the purview of dialectology, because they used to be overwhelmingly German-speaking. However, they have been Czech-speaking for almost 70 years now, following the post-World-War-II displacement of the German population and replacement with mostly Czechoslovak settlers. In exploring the possibility of local speech patterns which might have emerged and stabilized in these territories over the intervening years, the traditionally defined borderland regions are woefully inadequate as a unit of generalization, since they lump together geographically distant and sometimes even non-contiguous locales. To be fair, they were never intended as meaningful divisions for the purpose of investigating regional variation in Czech; their purpose was purely to circumscribe the territory

where such investigation was deemed possible or worthwhile. The dialectologists who delineated them only ever claimed ***hic sunt leones*** as far as Czech dialects are concerned. But if we ever want to make sense of what's happening there now, linguistically speaking, much like Polish dialectologists have begun attempting to make sense of the ***nowe dialekty mieszane*** (new mixed dialects, cf. e.g. Karaś 2010) of the formerly German territories in the north and west of modern-day Poland, we will have to look past these traditional boundaries and begin incorporating them into the broader story of regional variation in Czech.[8]

Similarly, constraining prosodic annotation to one of the existing abovementioned schemes runs the risk of throwing away variation that might well be worth exploring, but that these existing formalisms are not well-adapted to encode – perhaps because they were deemed non-central given the perspective adopted in the formalism, or because they are recent or emerging innovations. One such phenomenon might be the appearance of the high rise terminal (HRT), also known as uptalk, in Czech, possibly under influence from English (Zaepernicková & Havlík 2017: 55–6).

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

[8] To be clear: this should not be construed as criticism of my colleagues from the spoken corpora section at the CNC. I have been part of the team for almost a decade now, and could have pushed for a change in the direction I'm suggesting here at any time. The fact that I've so far failed to do so is entirely my own fault.

possibly elegant but quite probably also restrictive system of analysis.

To couch this in an analogy in terms of two well-known transcription systems for prosody: the prosodic annotation in a general-purpose spoken corpus should be more like INTSINT (cf. above), and less like ToBI (Beckman, Hirschberg & Shattuck-Hufnagel 2005). Using ToBI for annotation requires a relatively involved and intricate language-specific analysis in the generative tradition, which tries to come up with a model of intonation for the given language which is as sparse and elegant as possible, reflecting presumed underlying representations. Putting aside to what extent the assumptions and criteria of this paradigm are justified, it simply makes ToBI way too theory-laden to use as annotation approach in a general-purpose spoken corpus, and foist it by default upon every researcher who wants to use it. People interested in using ToBI for analyzing data from such a corpus are of course welcome to do so, but they must perform their own annotation, backed by their own theoretical decisions and tradeoffs, perhaps even leveraging the more descriptive prosodic annotation that comes with the corpus out of the box as a starting point, for reference or for quickly sifting through the data and identifying samples to annotate.

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

Dictionaries are like alphabets – extremely useful tools for working with language in normal life, but a potential distraction when engaged in the meta-endeavor of examining language, trying to tease it apart and figure out how it works. Alphabets sometimes make us forget that discrete phones or phonemes are an abstraction we impose on the sounds we make when we speak, an abstraction which is sometimes useful, sometimes less so (Port & Leary 2005; Ramscar & Port 2016). This is confirmed even by highly practical applications such as Automatic Speech Recognition (ASR) or speech generation: while the mainstream contenders opt for discrete inventories for modeling speech sounds, these are typically not isolated (mono)phones but triphones, to take into account coarticulatory effects.

Similarly with dictionaries, though it's perhaps less widely acknowledged. Dictionaries are of course enormously useful in everyday life, but the reverse of that is that they have become the dominant metaphor for what words are and how they

work, cemented in by Saussure's mnemonic triad of sign, signifier and signified. Intuitively, the signifier is the headword, the signified is the definition, and they are bound together in an indissociable sign, the full dictionary entry. For Saussure, this would of course only be a first crude approximation, intended for students to start wrapping their heads around the concept. What he ultimately had in mind was somewhat more sophisticated, with both signifier and signified existing only as elements of the linguistic system, defined purely by relations and contrasts with other elements in the system. But no matter, as we'll see below, such an extreme position is even more wrong and untenable than the popular approximation simmering at the back of all our minds.

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.

Lest I be accused of setting up a straw man just to be able to skewer it: Pelletier (2001: sec. 1) cites about a dozen linguists and philosophers who invoke this principle. In the strand of cognitive science which has roots in the generative grammar tradition, at least as far as language goes (via figures like Steven Pinker and Jerry Fodor), compositionality is one of the central tenets, the key to how the human mind processes and understands language. Gary Marcus, a former grad student of Pinker's, elaborated on this topic at length in his book ***The Algebraic Mind*** (Marcus 2001), where he criticized artificial intelligence systems (typically, neural networks) which avoid symbols and symbolic manipulation (centrally, composition), arguing that such systems can never, due to the resulting inherent limitations in their design, reach human levels of intelligence and ability to use language. Twenty years later, surveying the landscape of deep learning, which has seen tremendous in-paradigm advancements like GPT-2 (Radford et al. 2019), GPT-3 (Brown et al. 2020) or DALL-E (Ramesh et al. 2021) in recent years, Marcus is not impressed and holds his ground:

> The latest alleged triumph is that Google hinted in working paper for a new system called Imagen that they had made a key advance in one of the biggest outstanding problems in artificial intelligence: getting neural networks (the kind of AI that is currently popular) to understand ***compositionality***—understanding how sentences are put together out of their parts.

> The dirty secret in current AI is that for all the bloviating about how current systems have mastered human language, they are still really weak on compositionality, which most linguists (since Frege, over a century ago) would agree is at the very core of how language works.
>
> [...]
>
> For me, compositionality has always been the touchstone for AI; when I say AI has hit a wall, it's compositionality, more than anything else, that I have been talking about. (Marcus 2022)

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.[1] 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'. As for the

[1] Interestingly, Pelletier (2001: sec. 2) mentions a second tradition of discussing "Frege's Principle", where the principle in question is taken to mean something quite different: Baker & Hacker (1980: 258; quoted in Pelletier 2001: 90) formulate it as "the dictum 'a word has a meaning only in the context of a sentence'", going so far as to credit Frege with "destroying the grip of semantic atomism [on modern philosophy]". This is directly at odds with "Frege's Principle" *qua* compositionality: it's basically a different way to phrase the critiques we're currently examining.

proposed solutions to this conundrum, I will preferably borrow a paraphrase from someone involved in this general area of research, as I fear I would not be able to do them justice:

> The interpretation of so-called ***logical metonymy*** (e.g, ***The student begins the book***) has received an extensive attention in both psycholinguistic and linguistic research. The phenomenon is extremely problematic for traditional theories of compositionality (Asher, 2015) and is generally explained as a type clash between an event-selecting metonymic verb (e.g., ***begin***) and an entity-denoting nominal object (e.g., ***the book***), which triggers the recovery of a hidden event (e.g., ***reading***).
>
> [...] Thus, logical metonymy raises two major questions: i.) How is the hidden event recovered? ii.) What is the relationship between such mechanism and the increase in processing difficulty?
>
> One of the first accounts of the phenomenon dates back to the works of Pustejovsky (1995) and Jackendoff (1997), which assume that the covert event is retrieved from complex lexical entries consisting of rich knowledge structures (Pustejovsky's ***qualia roles***). For example, the representation of a noun like ***book*** includes telic properties (the purpose of the entity, e.g. ***read***) and agentive properties (the mode of creation of the entity, e.g. ***write***). The predicate-argument type mismatch triggers the retrieval of a covert event from the object noun qualia roles, thereby producing a semantic representation equivalent to ***begin to write the paper*** (see also the discussion in Traxler et al. (2002)). (emphases and bibliography references in the original, Chersoni, Lenci & Blache 2017)

This all sounds very sophisticated, but the basic answer is the same as that of the lexicographers: build (or posit in the mind) more (or at least, more sophisticated) dictionaries, cf. "complex lexical entries consisting of rich knowledge structures". While this particular angle comes from the generative tradition, even frameworks which are relatively critical of some of the basic tenets of generative grammar and have sought to overcome some of its limitations give answers which broadly reduce to this. I have specifically in mind CxG, which does not invoke a separate mental lexicon and can therefore sidestep any debates as to how rich the information is that is supposedly stored in there. Instead, it describes speakers' linguistic knowledge as construction inventories (organized in various ways), with the definition of each construction being able to refer to salient features cutting across all available levels of linguistic description, from phonology to pragmatics.

On this view, the building blocks work slightly differently: more complex constructions have slots where you can fit other constructions, and can in turn fit into even more complex constructions. "Traditional" compositionality says, words have meanings and are composed together into higher entities like sentences, which are ontologically a different kind of object, so where does the extra meaning come from? CxG answers, words and sentences are actually ontologically equivalent, they are both different types of constructions, and any construction in the hierarchy – in the edifice built from these building blocks – can contribute meaning to the whole. Not just the primitive ones (words), but even the complex ones with slots, which describe how we conventionally fit words together. But this is still a compositional approach, and the answer to the question ***How do we explain that the meaning of the whole seems to be more than the sum of its parts?*** is still: build (conceptualize) better (more sophisticated) dictionaries, which go beyond individual words – i.e. in this particular case, construction inventories.

So what is the correct answer to this question? I'll start with a few attempts at a short answer, which will probably sound like impenetrable zen koans at this point. Please come back to them later, after you've read more of the discussion and some examples, they will hopefully start making sense by then. The reason why I'm doing it this way is that I'm hoping it will make you experience first-hand how meaning actually works.

3.2 Meaning discrimination and information theory

Without further ado: 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"[2] 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

[2] Strictly speaking, Shannon's seminal 1948 paper, later reprinted in book form (Shannon & Weaver 1964), designates its topic as "mathematical theory of communication", but "information theory" ended up becoming the umbrella term for this and related areas of research.

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:

> As a starting place for this let us consider what factors are involved in communication; whether conducted by wire, direct speech, writing, or any other method. In the first place, there must be a group of physical symbols, such as words, dots and dashes or the like, which by general agreement convey certain meanings to the parties communicating. In any given communication the sender mentally selects a particular symbol and by some bodily motion, as of his vocal mechanism, causes the attention of the receiver to be directed to that particular symbol. 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. It does not, however, eliminate possibilities regarding the size of apples, and this further information may be conveyed by subsequent selections. (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. One part of the formulation is rather vague however, or potentially ambiguous: the claim that "words … convey … meanings". Since we are conditioned by the dictionary metaphor and compositionality, it is all too easy to slip into the erroneous interpretation that words ***carry*** meanings from speaker to hearer, from sender to receiver, and the overall meaning is somehow constructed out of these component units. I don't think this is what Hartley meant, because if you think it through, this is incompatible with the discriminative approach to meaning that he clearly lays out. I'll point out below, in the course of a short historical detour, formulations which I feel are closer in spirit to what he intended, quite surprisingly made by even earlier figures in the history of philosophy and linguistics. Nevertheless, this is exactly the alluring conceptual shortcut that MacKay, and by way of him also Jakobson, took – the dictionary trap that they fell into:

> GIT [General Information Theory] was chiefly elaborated with the purpose of giving Shannon's information definition a place within disciplines outside of engineering. For that reason MacKay also defined the subject matter of GIT broadly as the field studying "the making of representations" (MacKay 1952: 42). Importantly, he further split GIT into two sub-disciplines, the first of which he called Communication Theory (hereafter CT) because it solely deals with processes on prefabricated representations (MacKay 1952). According to Jakobson linguistics now had to be viewed as an instance of Communication Theory. (Van de Walle 2008: 114)

If you squint hard enough through the fresh coat of paint, then this is the familiar dictionary + composition model all over again. "Representations" are akin to

dictionary entries, separately existing entities (they are first made and subsequently used during communication) which rely on the integration of signifier and signified, a form alongside a unit of meaning, as the secret sauce which makes communication work. Not heeding Shannon's explicit warning not to conflate information and meaning, this interpretation sees the encoding of the message on the sender's side as wrapping up a brick of meaning in a nice little box, then sending it over through the channel, and then decoding consists in the receiver unwrapping the box, recovering the meaning brick and sticking it into the appropriate place in the brick tower of meaning he or she has been building. Such a conceptualization is open to 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*[3]:

> Interestingly, Jakobson seemed to embrace information theory with its notions of code, message, sender and receiver, as a way of returning the act of communication to center stage in linguistics. Linguists, felt Jakobson, had fallen into the habit of fetishizing the forms of language, without considering what those forms are actually used for – what they actually mean. This approach to linguistics, which actively banished meaning as a valid topic of linguistic research, was particularly characteristic of the school that formed around Leonard Bloomfield in the first half of the 20th c. [...]
>
> 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

[3]Available online at https://hiphilangsci.net/.

> is encoded and decoded, and noise may interfere with the message, but there is always a single message that is in principle recoverable. Shannon and Weaver, in their 1949 book,[4] insisted that information theory was not concerned with meaning. However, Weaver did allow for the tantalizing possibility that information theory at the very least laid the necessary groundwork for the study of meaning, and that meaning might very soon be within its grasp.
>
> But this story of senders, receivers and codes is a long way from the approach to meaning we met in the previous episodes. Such figures as Wegener, Lady Welby, Firth and Malinowski all treated meaning in a way that we could broadly characterize as hermeneutic. That's to say, they emphasize that exchange of words and other meaningful symbols is an active process both for the producer of the symbols, and those trying to interpret them. There are constraints on interpretation – it's not the case that "anything goes" – but there is room for genuine ambiguity, and different interpreters might legitimately arrive at different meanings. Indeed, the meaning that arises in a particular situation might even surprise the producer of the symbols. (McElvenny 2022 cca 25:00–28:00)

On the face of it, this seems like a serious objection, and a critical flaw which makes information theory unsuitable as a framework for couching the tricky and elusive finer points of human communication. However, the problem lies not in information theory, it lies in equating information with meaning, which Shannon was warning against all along, and thus falling into the dictionary trap. The ap-

[4] Sic; while Shannon and Weaver's original articles are indeed from the late 1940s, the book collecting them was only published in 1964.

propriate response should therefore be familiar by now: reject the premise of a dictionary-based model of meaning, reject the premise of compositionality. Strictly speaking, words do not somehow possess, or carry, meaning, even though that's how we typically think and talk about them.[5] Meaning does not travel across the communication channel, whatever the channel is; only messages travel. Meaning stays put in your head, my head, everyone's head; what's sent over the channel is information, which helps you *discriminate* meaning. This is borderline compatible with saying that words "convey" meaning, as Hartley puts it, because while 'convey' can be a synonym to 'carry', it also has more abstract connotations and can be interpreted as meaning 'communicate', 'impart' or even 'evoke'. Still, as I've noted, given the sheer inertia of the dictionary metaphor in our minds, the formulation is dangerously equivocal, because it's likely to be read as "words carry meaning".

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. Since the assumption is not only not a core part of information theory, but actually antithetical to it, according to Shannon as its seminal theorist, it follows that without it, information theory, far from being "a long way from" the "hermeneutic" approach of Wegener, Firth and others, remains at the very least compatible with it. Even further, to my mind, it's actually a perfectly natural fit.

[5] Including myself, in contexts where the distinction doesn't matter and going against established convention would just make it harder to make myself understood.

One thing I want to make clear is that in accounting for how different meanings can possibly emerge in a discriminative model, I'm not trying to chalk it up to top-down processes in the brain messing with bottom-up raw sensory input. This of course happens routinely, because top-down prediction is crucial for processing of sensory input to be efficient and close to real-time. Most of the time, if the predictions are wrong, a conflict occurs between prediction and raw sensory data, and mismatch error is propagated up the neural hierarchy until the prediction is adjusted and the conflict is resolved. Sometimes however, the raw sensory signal is so weak, or the prediction is so strong, that it overrides reality, so to speak, and we hear what we think we should be hearing, colloquially speaking, instead of what is actually being said. In such situations, differences arise already at the step where the message is decoded, i.e. even before it is interpreted, before it is used to discriminate meaning.[6]

A toy example: Alice says *A B D*. Bob is close by, and correctly hears *A B D*. Carol is in the next room, the sound of Alice's voice is muffled, and she strongly expects *C* to come after *B*, so she mis-hears *A B C*. In McElvenny's terms, the single definite message here is Alice's original *A B D*; Bob succeeds in recovering it and Carol fails. However, this type of message-level (or information-level) miscommunication is perfectly recoverable: Alice can repeat the message louder, she can go into Carol's room (or Carol in Alice's), she can even write the message down in order to make really sure the correct symbols get across. Getting Alice, Bob and Carol to agree that Alice originally said *A B D* is not only possible in principle, it is also very much achievable in practice.

[6] Positing 'decoding' and 'interpretation' as two separate steps that happen in sequence is a simplification to allow us to discuss two different types or sources of ambiguity. I'm not claiming this is how it's implemented in the brain; rather to the contrary, I believe that in reality, these two processes are interleaved. What has already been interpreted contributes to shaping future predictions on what might be said next, which in turn affects what is actually decoded.

By contrast, getting Alice, Bob and Carol to agree on what Alice *meant* by *A B D* is *impossible* in principle, but *asymptotically approachable* in practice. Why impossible? Because you can never really get into another person's head and retrieve his or her exact perspective, including all the connotations and shades of meaning a certain message might evoke. Why asymptotically approachable? Because when it matters, discrimination of meaning is heavily constrained by the real world and our actions in it.[7] It's not trivial, even with people you know well and have a good model of what's inside their heads – in some cases, perhaps especially with those, because it can lead to a false sense of confidence, where you think you already know what they think and discount evidence to the contrary. But in principle, you can keep getting ever closer to intersubjective alignment. This is similar to how probabilities of 0 and 1 don't really make sense and break probabilistic reasoning and computations; in practice, all real-world probabilities are somewhere in between (see e.g. Yudkowsky 2015: chap. 55 0 And 1 Are Not Probabilities).

In summary, 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.

An analogy for the technically-minded, to show that this line of thinking is also easily applicable to the technical fields for the use of which information theory was initially elaborated: in a programming context, the source code / bytecode / machine code of a program is the message. It's always the same. But depending on

[7] A recent example: if a neighboring nation calls itself a friend and a brother, but keeps encroaching upon your territorial sovereignty, sooner or later you'll realize that what they mean by 'friend' and 'brother' is not what you think they meant.

how, where and when you run the program – what hardware is available, what security vulnerabilities it's susceptible to, what interpreter or OS version you're using, etc. – its behavior may differ quite substantially, even when everything that would conventionally be considered as input to the program is kept constant. That behavior is the program's meaning on that particular system. As with natural language, the meaning is ultimately not in the code, but in actual physical systems executing it, whether they be machines, or people doing so in their heads. Importantly, this shows that the account I've presented is not somehow a hack that needs to be bolted on information theory to make it work with natural language; it works equally well in other areas of application of the theory.

In the interest of relative brevity, I'm omitting evidence and arguments in favor of a discriminative theory of meaning which are not exactly germane to the topic at hand, and which would require non-trivial digressions to fully lay out. For instance, Ramscar (2019) deals extensively with the fact that unlike artificial information systems, the codes underlying human languages are neither fixed (as McElvenny also observes) nor pre-established in advance. He argues that the statistical structure of linguistically meaningful inventories of competing items is such that speakers converge towards very similar mental linguistic systems, even though the code is in principle infinite and ever-changing and no one speaker has full command of it, the key point that the rank×frequency relationship in these inventories is exponential[8] and therefore memoryless (see also Linke & Ramscar 2020; Ramscar 2020). Beyond that, experiments in child language acquisition have shown that language learning follows discriminative principles, as modeled by the Rescorla-Wagner (Rescorla

[8]Note that the traditionally observed "Zipfian" (Zipf 1949) relationship between rank and frequency in language is not exponential but a power law. Ramscar claims this is the result of aggregating units which do not actually compete with each other, which do not form a linguistically meaningful inventory.

& Wagner 1972) learning rule (Ramscar & Yarlett 2007; St. Clair, Monaghan & Ramscar 2009; Ramscar et al. 2010; Ramscar, Dye & Klein 2013; Ramscar, Dye & McCauley 2013). Psycholinguistic evidence is also available from studies on lexical processing response times in adults (Baayen 2010; Baayen et al. 2011; Baayen, Hendrix & Ramscar 2013; Milin et al. 2017), the paradigm has been successfully used for computational modeling of lexicon and morphology (Baayen, Chuang & Blevins 2018; Baayen et al. 2018; Chuang & Baayen 2021), and even attempts to bypass phonemic segmentation in models of auditory perception (Baayen et al. 2016; Arnold et al. 2017).

3.3 First-hand evidence: Skin in the game

Going back to the example Hartley uses to demonstrate the discriminative nature of communication – "Apples are red" – it may sound unconvincing and made up, as such – made up – examples tend to do. You could argue you'd be able to give a plausible sounding formulation of the same process in a compositional framework, and it would feel equally compelling at first blush. Or conversely, with the discriminative lens having become my default perspective on communication, I sometimes wonder whether I might be suffering from confirmation bias. Do I see language phenomena in a discriminative light just because I've come to expect to see them that way, not because it's the most fitting explanation? This is why I've started collecting personal experiences where the discriminative interpretation stands out as particularly enlightening. In other words, examples that are the polar opposite of the blandness of "Apples are red" and suchlike, which seem like they could be used either way, just with a different analysis slapped onto them.

How valid are personal anecdotes from a scientific point of view? For one thing,

certainly more valid than entirely made up examples, which are nevertheless commonly both proferred and accepted in linguistics. For another, Grieve (2021: 1) points out that language is "an inextricably social phenomenon", which means "linguists should not be surprised when experimental results fail to replicate", and should also acknowledge "the scienfitic value of observational methods". I heartily concur, adding that observational methods can include casuistry and even introspection. While not quantitative, they can still be empirical, provided they refrain from speculation and stick to relating facts. This is obviously a fine line to tread and places heavy demands on personal rigor and honesty; readers should approach such observations with a healthy critical spirit, but dismissing them out of hand seems unwarranted to me.

If you squint, you can view it as corpus data that just never happened to be collected in a corpus, approached in a conversation-analytic mindset. It's also much richer in detail and nuance than typical corpus data, compared to what audio or even video recordings (in the case of spoken corpora) can offer, where you typically have very little information about any of the parties involved, or the context beyond what's overtly discussed. With interactions you've personally witnessed or taken part in, you know at least yourself pretty well, and have access to your own inner experience, but you also typically know the other participants much better than the anonymous speakers in corpora. Consequently, you have a more elaborate predictive model of their behavior, which can in turn allow you to pick up on even relatively innocuous cues, enhancing your perception of the entire situation (perception is in large measure prediction, as we've seen).[9]

[9]There are downsides too, of course. For one thing, a strong predictive model has a lot of inertia, it's harder to update. In other words, when you know someone well, you can fall into the trap of letting your model override sensory evidence that contradicts it, but is too weak to warrant changing your prediction. But perhaps the most conspicuous trap is confirmation bias – our tendency to recall with great ease experiences which confirm our current beliefs, while burying

With such matters of principle out of the way, I'll start by giving two personal anecdotes, then move on to discuss the example of an online community which, from a starting point of valuing clear communication about complex topics and an openness to change one's mind, has naturally arrived at attitudes towards communication which are very close to what I mean by discriminative here. To my knowledge, the term itself does not feature prominently in their discussions on the topic, or at least I've never seen it explicitly mentioned, which leads me to conclude they developed these ideas independently, in a separate tradition.

Early on during my tenure as a spoken corpus transcription coordinator at the CNC, I came across what seemed to me like a clear typo in a transcript of a speaker from Silesia (see Figure 2.3), the easternmost part of the country with whose speech I'm less well acquainted than with that of Prague, my birthplace and hometown. While agreeing with the phonetic transcript layer that the pronunciation of the target word was [kura], phonetically speaking, I thought that transcribing it as ***kura*** on the basic transcript layer was an error, as I'd never heard of such a word before, and amended it instead to ***kurva***, a rather strong expletive meaning 'fuck', lit. 'whore'. I want to emphasize I did not consider ***kura*** as an alternative, and then discard it; from my perspective, this alternative did not exist at all. However, when sharing feedback with the original transcriber, she was horrified: the speaker was a relative of hers and would never use such crude language in this context; he actually did say ***kura***, which she explained to me is a common, and crucially much tamer, expletive in Silesia, quite distinct from ***kurva***.

So what did the speaker ***actually*** say? In a compositional framework, where meanings are properties of words, this is an inextricable conundrum. I'd ***heard*** the word correctly as [kura], not the canonical pronunciation [kurva], so in terms of

those which disconfirm them.

the foregoing discussion, there was no ambiguity at the decoding stage to chalk it up to. How can a single form have two different meanings at the same time?

In a discriminative framework, the answer is straightforward: words in and of themselves do not *possess* any meaning at any time, meaning is in the mind. In my case, ***kurva*** was the only available option, so that's what I ended up with. On the other hand, in the case of the speaker and transcriber, who distinguish (discriminate) between ***kurva*** and ***kura*** in their Silesian idiolects, it was clearly the latter. I genuinely thought the speaker had used crude language; they genuinely thought he hadn't.

The second anecdote revolves around a personal name, which already poses a challenge to compositional approaches: "some of the biggest problems posed by the idea of meaning compositionality have been encountered in relation to personal names (Frege, 1892; Russell, 1919; Searle, 1971; Donnellan, 1972; Burge, 1974; McDowell, 1977; Boersema, 2000)" (Ramscar 2019: 11, references in the original). In the course of discussing children's literature, a friend of mine recommended a series of books by one Ladislav Špaček. I distinctly remember briefly considering the name before concluding: doesn't ring a bell. She then went on to say that two of the core topics were ethics and etiquette, which made me realize, hang on, I ***do*** know who Ladislav Špaček is after all! He's this self-professed etiquette expert who sometimes comes up in the media.

From the point of view of how I subjectively experienced it, this episode was even more startling than the ***kura*** vs. ***kurva*** misunderstanding: while ***kura*** was a new word to me at the time,[10] I've known who Ladislav Špaček is for about as long as I can remember – he was a relatively prominent figure on the news during my childhood in the 1990s, as spokesman of president Václav Havel. But for a few

[10] And I encounter it so rarely that when I hear it and don't think too hard about it, it's still liable to trigger ***kurva*** instead.

moments there, it felt as if I'd never heard the name. What happened?

Again, it's about what options you start with at the outset, to what degree and in what ways the metaphorical lump of stone is pre-carved by expectations based on your previous experiences in general, as well as the specific interactional context, including the conversation so far. In my case, the set of people I know in a general context includes Ladislav Špaček, as well as many other people, including authors of children's literature like Astrid Lindgren and J. K. Rowling. However, I'd had previously no idea that Špaček also wrote children's books. So in the context of children's authors, he simply didn't exist in my mind, the lump of people to discriminate among had been pre-carved to exclude him – I'd been contextually blinded to him. I want to make it clear that my initial reaction was ***not*** "Ladislav Špaček? I know an etiquette expert by that name, but no children's author, that must be someone else"; it truly was "this is not a name I remember hearing before, at all". Then hearing additional information about etiquette made me backtrack the discrimination process and restart it with a broader initial set of people, not only those I associate with being children's authors, and suddenly of course I knew who he was. On his own, Špaček was too weak as evidence against the restrictions imposed by the children's literature context, and didn't trigger the backtracking (after all, it's not an especially unusual name), but Špaček + etiquette tipped the balance.[11]

Now, you might argue that these experiences are highly subjective and that I'm biased to favor a discriminative interpretation of communication, so while I may not technically be lying, much of what I'm reporting may actually be in the eye of the beholder (i.e. myself), so to speak, rather than a reliably generalizable property of the phenomenon itself. So let's take a look at a community of people

[11] I won't attempt a compositional analysis in this case and leave it as a (futile) exercise to the reader.

who occasionally write about similar experiences, but from a different perspective, as their primary concern is not linguistic theory, discriminative or otherwise, but clear communication.

The community in question is the so-called rationalist community, whose most iconic online gathering space is probably the server *LessWrong.com*. Nowadays, the term "rationalist" has acquired negative connotations in some circles, evoking people who are self-confident verging on arrogant, never change their mind, and automatically expect everyone else to accept rational arguments, no matter how they're phrased. By contrast, the self-described rationalists of *LessWrong* see this as a failure mode of the enterprise of trying to be a rationalist. This is enshrined in the name of server itself, which members of the community elucidate thusly: "We might never attain perfect understanding of the world, but we can at least strive to become less and less wrong each day."[12]

From this, it should be clear that epistemic humility is a key virtue for rationalists. Humility does not imply weakness or endless relativism; be confident where confidence is warranted given available evidence, and do your best to convince other people to adopt your viewpoint, especially if the stakes are high.[13] But also remain open to genuinely consider revising your stance should compelling evidence to the contrary emerge. Rationalists consider changing one's *own* mind as perhaps the most important skill for an aspiring rationalist; without it, you can't take a single step on the path to being less wrong. Eliezer Yudkowsky, an important figure within the community from its early days, devotes an entire book section to it (2015),

[12] See https://www.lesswrong.com/about.

[13] The rationalist community has a high degree of overlap with the effective altruist community, which can be seen as its ethical branch. At the risk of oversimplifying: effective altruism applies the methods of rationality to helping other people as effectively as possible, which includes trying to convince others to do so as well. If "as effectively as possible" sounds problematic or naive to you, you're not alone. This aspect has sparked numerous controversies with good arguments on either side; I won't go into them here.

entitled *How to Actually Change Your Mind*.

The ability to change one's own mind, as well as to change the mind of others, crucially hinges on effective communication. As speakers, this involves soft skills like emotional intelligence[14] (not typically associated with the negative rationalist stereotype I mentioned), an ability to adapt and attune oneself to the needs of one's communication partners, as well as intellectual skills like being able to estimate what can and cannot be conveyed given a certain amount of time and shared background. As listeners, this mostly involves being aware of the many cognitive biases humans are endowed with at birth, as a result of our evolutionary history, and fighting tooth and nail to counteract them when they risk clouding our judgment.

So the rationalists are a community of people who try really hard to communicate clearly about complex topics, who truly care about understanding others, as well as being understood by them, correctly. My contention is that they've developed a tradition and strategies around this which is very much reminiscent of what I've been talking about in this chapter, without necessarily explicitly reflecting on the discriminative nature of meaning. The intuition behind the name *LessWrong* should already remind you of the asymptotic approach to meaning I outlined above. But I should probably let them speak for themselves, and let you be the judge as to how close the views are to those I'm advocating. Here are a few quotes from a foundational series of blog posts written by Eliezer Yudkowsky, often referred to simply as "the sequences". They became very influential within the community and frequently quoted; they were later edited into a book, *Rationality: From AI to Zombies* (2015), which is where I'm quoting from for easier reference.

On the illusion of transparency of our own thoughts:

[14] "I'm not saying that I think we should be apolitical, or even that we should adopt Wikipedia's ideal of the Neutral Point of View. But try to resist getting in those good, solid digs if you can possibly avoid it." (Yudkowsky 2015: 256)

> [T]he *illusion of transparency*: We always know what we mean by our words, and so we expect others to know it too. Reading our own writing, the intended interpretation falls easily into place, guided by our knowledge of what we really meant. It's hard to empathize with someone who must interpret blindly, guided only by the words.
>
> [...]
>
> "The goose hangs high" is an archaic English idiom that has passed out of use in modern language. Keysar and Bly told one group of subjects that "the goose hangs high" meant that the future looks good; another group of subjects learned that "the goose hangs high" meant the future looks gloomy. Subjects were then asked which of these two meanings an *uninformed* listener would be more likely to attribute to the idiom. Each group thought that listeners would perceive the meaning presented as "standard."
>
> [...]
>
> As Keysar and Barr note, two days before Germany's attack on Poland, Chamberlain sent a letter intended to make it clear that Britain would fight if any invasion occurred. The letter, phrased in polite diplomatese, was heard by Hitler as conciliatory—and the tanks rolled.
>
> Be not too quick to blame those who misunderstand your perfectly clear sentences, spoken or written. Chances are, your words are more ambiguous than you think. (Yudkowsky 2015: 34–6, emphases in the original)

On the mismatch between the ancestral environment to which our cognitive

skills are originally adapted, and the environment into which a modern human being is thrust upon birth:

> ***Homo sapiens***'s environment of evolutionary adaptedness (a.k.a. EEA or "ancestral environment") consisted of hunter-gatherer bands of at most 200 people, with no writing. All inherited knowledge was passed down by speech and memory.
>
> In a world like that, all background knowledge is universal knowledge. All information not strictly private is public, period.
>
> [...]
>
> When you explain things in an ancestral environment, you almost ***never*** have to explain your concepts. At most you have to explain ***one*** new concept, not two or more simultaneously.
>
> In the ancestral environment there were no abstract disciplines with vast bodies of carefully gathered evidence generalized into elegant theories transmitted by written books whose conclusions are ***a hundred inferential steps*** removed from universally shared background premises.
>
> In the ancestral environment, anyone who says something with no obvious support is a liar or an idiot. You're not likely to think, "Hey, maybe this person has well-supported background knowledge that no one in my band has even heard of," because it was a reliable invariant of the ancestral environment that this didn't happen.
>
> Conversely, if you say something blatantly obvious and the other person doesn't see it, ***they're*** the idiot, or they're being deliberately

obstinate to annoy you.

And to top it off, if someone says something with no obvious support and *expects* you to believe it—acting all indignant when you don't—then they must be *crazy*.

Combined with the illusion of transparency and self-anchoring, I think this explains a *lot* about the legendary difficulty most scientists have in communicating with a lay audience—or even communicating with scientists from other disciplines.[15]

[...]

A clear argument has to lay out an inferential *pathway*, starting from what the audience *already knows or accepts*. If you don't recurse far enough, you're just talking to yourself.

[...]

Oh, and you'd better not drop any hints that *you* think you're working a dozen inferential steps away from what the audience knows, or that *you* think you have special background knowledge not available to them. The audience doesn't know anything about an evolutionary-psychological argument for a cognitive bias to underestimate inferential distances leading to traffic jams in communication. They'll just think you're condescending.

And if you think you can explain the concept of "systematically underestimated inferential distances" briefly, in just a few words, I've got

[15] Or even communicating with scientists from the same discipline, but with a different theoretical or methodological background.

> some sad news for you… (Yudkowsky 2015: 37–9, emphases in the original)

Last but not least, on where meaning resides:

> Words do not have intrinsic definitions. If I hear the syllables "bea-ver" and think of a large rodent, that is a fact about my own state of mind, not a fact about the syllables "bea-ver." (Yudkowsky 2015: 121)

I expect especially this last one to sound very familiar by now. I intentionally selected relatively long-form citations to provide as much context as possible, and as I've said, these ideas have become quite influential in the rationalist community: people quite frequently refer to "the sequences" in their own blog posts or discussions, as a short-hand to these notions. I will not demonstrate this exhaustively, but I will provide one last representative example from an author other than Yudkowsky. The following quote is from a post on *LessWrong* by Duncan Sabien, entitled *Ruling Out Everything Else*, which is about as discriminative as you can get without actually saying the word "discriminative": as we've seen, discrimination is literally the process of ruling out possibilities.

> Clear communication is difficult. Most people, including many of those with thoughts genuinely worth sharing, are not especially good at it.
>
> I am only sometimes good at it, but a major piece of what makes me sometimes good at it is described below in concrete and straightforward terms.
>
> The short version of the thing is "rule out everything you didn't mean."

> That phrase by itself could imply a lot of different things, though, many of which I do not intend. The rest of this essay, therefore, is me *ruling out everything I didn't mean* by the phrase "rule out everything you didn't mean."
>
> [...]
>
> - Notice a concept you wish to communicate.
> - Form phrases and sentences which accurately match the concept as it lives in your mind.
> - Notice the specific ways in which those phrases and sentences will mislead your audience/reliably trigger predictable confusions. [...]
> - Pre-empt the confusion by ruling out those misunderstandings which are some weighted combination of "most likely," "most common," "most serious," and "most charged."
>
> (Sabien 2021)

I quote from this post as aligned with my thinking, in spite of the fact that one of the headings reads ***Words have meaning (but what is it?)***, which at first blush seems to be completely opposite to what I'm saying, or what Yudkowsky said in that last quote of his, for that matter. But in the spirit of ruling out what one didn't mean by what one said, I would argue this should not in fact be construed as a claim on where meaning is located. What Sabien meant by this is that words are used ***conventionally***, and if you want to be understood, you need to try and abide by these conventions. So in discriminative parlance, we would say that there are conventions as to which words discriminate which meanings. In this sense, the claim that "words have meaning" is absolutely true, and it's also a much more natural sounding way

to put it than my rather technical and abstruse paraphrase. In a way, it's very apt that even a sentence like "words have meaning", quite simple and straightforward on the face of it, can act as both congruent and incongruent with a discriminative approach to meaning, depending on which meaning it happens to discriminate in your mind. Still, I would tend to avoid it, because it feels like dancing dangerously close to the edge of the dictionary trap. So why doesn't Sabien?

Because his goal is different from mine. As a linguist, I'm concerned with how language generally, or meaning specifically, works in the abstract, so as to provide a firm foothold for the study of language. By contrast, Sabien is laying out a practical strategy for minimizing the likelihood of being misunderstood in communication. His practical advice – try to think about how your formulations could be misinterpreted, and prevent the confusion by explicitly ruling out these possibilities – is not hindered in any way if the reader happens to understand "words have meaning" in the wrong way. The text is not about what words are or how they work, it's about what you should do if you're trying to get a complicated point across. More generally, the example of the rationalist community shows how thinking about communication from a practical point of view, driven by practical needs, steers you in the right direction, even if you don't necessarily have (or care about) the right underlying theory.

Unfortunately, a corollary of this is that linguists often have the wrong incentives. When investigating language, what's foremost in our minds is usually not clear communication (in the service of effecting positive change in the real world), but in the best case scenario, dissecting language, figuring out how it works. That's a worthy goal, do not mistake me, but it can lead to theory for the sake of theory, which fails to account for evidence that would be blatantly obvious in a more practical mindset. In the worst case, our investigations are shaped by the need to

play status games, as is common across institutionalized science, or indeed any social environment where status hierarchies play a role. In status games, different sorts of considerations altogether apply: strategic conformity and deference to tradition, or conversely, status-seeking disruption and innovation, signaling, boom-and-bust cycles, etc.

A possibly useful analogy: Nassim Taleb has devoted an entire book to the concept of ***skin in the game*** (Taleb 2018). With a background in finance, an example distinction that Taleb often draws is between economists (who only theorize about how markets work), fund managers (who gamble other people's money for a share of the winnings, but don't participate in any losses) and traders who invest their own money. According to Taleb, only the third category has the correct set of incentives, both reaping the reward, but also shouldering the accompanying risks and possible losses. Only a full, balanced set of incentives, both positive and negative – in other words, only truly having skin in the game – leads to behavior and beliefs that other people can consider trustworthy, or at least the best possible attempt given an individual's skills and available evidence.

In this analogy, we as linguists are unfortunately the economists.[16] In case you're wondering whether Taleb meant for the analogy to stretch beyond finance – in the introduction to the book, he's quite clear that the requirement of skin in the game applies broadly to "uncertainty and the reliability of knowledge (both practical and scientific, assuming there is a difference), or in less polite words bull***t detection" (Taleb 2018: 3). The yardstick we measure ourselves as linguists with is overwhelmingly not whether our theories tell us how to use language better or more effectively. This has undoubtedly been a contributing factor in Fred Jelinek's famous rule of thumb: "Every time we fire a phonetician/linguist, the performance of our system

[16] Who are the fund managers? Wholesale purveyors of questionable linguistic advice like Strunk and White?

goes up" (Moore 2005: 117). Perhaps if we pay more attention to skin in the game in the future, we can still overturn Jelinek's dictum.

3.4 Practical consequences for real-life communication

While I used the rationalist community primarily as circumstantial evidence in a theoretical dispute, namely in favor of a discriminative theory of meaning, I still think their practical insights are far-reaching and should not be taken lightly. Realizing that meaning works in a discriminative fashion allows us not only to better understand how we *do* communicate, but also how we *ought to* communicate, both on the sending side (speaking/writing/devising arguments) and the receiving one (listening/reading/processing others' arguments).

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.

Failing that, you can take a page out of Plato's playbook and opt for a genre which at least *pretends* it's dialogue. Using texts by the pseudonymous Scott Alexander, another prolific and popular writer in the rationalist community, as examples, this is a surprisingly common strategy in the rationalist community: either explicitly structuring texts in a question and answer format (often labeling them as FAQs, i.e. frequently asked questions, see e.g. Alexander 2016), or sometimes even outright

Platonic dialogues (e.g. Alexander 2014), or at least engaging in some kind of less formalized back-and-forth with a hypothetical reader. Alternatively, if a predominantly monologic style is what comes naturally to you in writing, don't be afraid of liberally quoting other authors writing on the same topic. It's a chance to let other people speak, which is always a good idea: it's a fair bet that with some members of your audience, other perspectives and wording choices than your own will resonate better.

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. Prefer rewording from various perspectives, seasoned with copious examples, try to anticipate possible misunderstandings (cf. Sabien above), rather than spending time and energy on refining concise formulations until they feel like they perfectly capture what you're trying to say. They feel like that *to you* because *you already know* what you're trying to say, so it takes very little input to discriminate that particular idea from the other ideas in your head. But from the perspective of someone who does *not* already have that idea in their head (which may include yourself, a few months or years down the road), there's precious little to grasp onto.

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.

These may sound commonsensical to the point of seeming banal, but speaking from personal experience, even though I know all this on a conscious level, I still struggle when trying to apply it in practice, both with respect to reading (I routinely get stuck on a single sentence for minutes on end) and writing, where I often tend to a dense and convoluted style, trying to say a lot in as few words as possible, skimping on concrete examples, spending instead a lot of time on condensed, abstract thoughts and getting them just right – a futile endeavor, as we've seen. One reason is undoubtedly having read a lot of material written in the same vein, and conforming to the style, but another is that saying a lot in few words, and just the right ones, feels intrinsically rewarding. However, one should consider that with each condensation, not only your text shrinks, your readership does too (cf. the quotes by Yudkowsky on the illusion of transparency and the necessity to lay out a clear inferential path).

Still, whenever I write, I try to break free of these propensities, and practice what I preach, even at the cost of occasionally departing from established conventions in academic writing, which may be especially obvious in the present chapter. To the degree that they are an obstacle to effective communication, I firmly believe that it's the conventions that need to change.

3.5 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.[17] 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. But for Leibniz, for instance, ideas are universal to the point of being computable: if the primordial meanings of words can be recovered through etymology, an engine can be constructed that will reason by applying algebraic operations to them (notice the similarity to what proponents of compositionality like Gary Marcus still claim today).

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

[17] I'm indebted to James McElvenny's previously mentioned podcast series, ***History and Philosophy of the Language Sciences***, for making me discover or rediscover some of these associations, even though I do not always adopt his interpretation of the facts.

idea of linguistic relativism[18] 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). Carefully circumscribing your subject area noticeably reduces the attack surface, which is a good thing if you don't like being caught off-guard. Also, as John Joseph takes care to point out in the quote above, Saussure did not altogether dismiss these alternative perspectives as futile pursuits, he just left them to specialists in other areas. And credit where credit is due: much can be achieved even with such a restricted mindset, as demonstrated by his pioneering work in the reconstruction of the Proto-Indo-European vowel system. But when faced with fundamental questions about the nature of language and meaning, I would argue leaving out the

[18] By contrast, roughly half a century after Saussure's death, Noam Chomsky bought the idea of a "universal grammar" from the Port-Royalists wholesale.

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.

Saussure clearly understood so much about language in some ways, and yet so little in others. His writing process, as characterized by John Joseph, is more or less the antithesis of the recommendations laid out in the previous section:

> Saussure had been trying and failing to write books about big methodological questions in the study of languages since his early 20s. The problem was that he was a perfectionist, determined that every word from his pen had to be precisely the right word – hence the thousands of draft manuscript pages in his archives that lay unpublished until recent years, in which the same thought is often recomposed ten, twenty times, then scratched through and abandoned. (Joseph & McElvenny 2022: 44)

Is it then any wonder that the one book that made him truly famous, the *Cours de linguistique générale*, was not actually written by him, but emerged in an iterative lecturing process, in dialogue with his students and posthumous editors? The lecture setting forced him to let go and say *something*, the best he currently had, each time he gave the course, see how it stuck, and perhaps try a bit differently next time. *This* is the essence of a successful communicative strategy, not agonizing over *le mot juste* for thousands upon thousands of pages of drafts.

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. This is assuming a form of naive realism, as per Yudkowsky (2015: 187–8): "[O]ccasionally I believe strongly that something is going to happen, and then something else happens instead. I need a name for whatever-it-is that determines my experimental results, so I call it 'reality'." In the context of communication: others' utterances lead us to make predictions about the future; when these predictions fail, we are forced to re-evaluate and re-adjust our models.

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; every separate item of which it is composed is, in its present form—for we are not yet ready for a discussion of the ultimate origin of human speech—the product of a series of changes, effected by the will and consent of men, working themselves out under historical conditions, and conditions of man's nature, and by the impulse of motives, which are, in the main, distinctly traceable, and form a legitimate subject of scientific investigation. (Whitney 1884: 48)

Saussure wanted to work with the conventions in the abstract, and that's fine up to a point, but there are questions which can only be answered if we bring those who "legislate" the conventions into the picture.

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.[19] 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:

> [N]ous verrons que nous faisons honneur au langage d'une quantité de notions et d'idées qu'il passe sous silence, et qu'en réalité nous suppléons les rapports que nous croyons qu'il exprime. J'ajoute que

[19] In the spirit of "ruling out", Whitney's position is perhaps best defined by who he had bitter polemics with. With respect to 19th century German thought on language, he vocally disagreed with the excesses of both Schleicher's exaggerated physicalism (Whitney 1873a) and Steinthal's highly speculative *Völkerpsychologie* (Whitney 1873b).

> c'est parce que le langage laisse une part énorme au sous-entendu, qu'il est capable de se prêter au progrès de la pensée humaine.[20] (Bréal 1868: 9)

Nerlich (1990: 99) gives an extensive summary of the article, which makes it even clearer:

> Wordforms give no direct access to meaning, they only give meagre hints, or minimal instructions, on the basis of which our intelligence, our mind, must construct meaning, ***make*** sense. That this is so should not be regarded as a shortcoming of language. On the contrary, if words represented exactly what they mean or refer to, as in some scientific nomenclatures, language in the normal sense would die, would no longer be usable, it would lose its function. Linguistic signs have to be vague and flexible for their users, so that they can be adapted to the wide variety of thoughts the users wish to express. This also means that linguistic signs are not created once and for all; they are constantly recreated and changed by those who use them. (emphasis in the original)

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). After all, having one's legacy oversimplified, or even adulterated, is often the penalty for success. We saw it with Frege who, according to Pelletier (2001), possibly didn't

[20] My translation: "We shall see that we attribute to language many notions and ideas that it is silent about, and that in reality, we supplement the relations we think it expresses. Additionally, I claim that it is precisely by leaving so much implicit and unsaid that language can serve the progress of human thought."

believe in anything close to the modern Frege's Principle *qua* compositionality; quite opposed to such atomistic views, an alternative tradition credits him with a holistic approach, where the meaning of words should always be examined within the context of a sentence (Pelletier 2001: sec. 2).[21]

3.6 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. Much like I did earlier in the case of the quote from Hartley (1928), Stewart takes issue with the common turn of phrase that meaning is "conveyed", because it obscures the reality of how language actually works:

> [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

[21] Do not mistake this as pining for a universe where Frege didn't get misinterpreted (assuming that is indeed what happened). The dictionary trap feels so intuitive and alluring that if not for Frege, we would have certainly come up with a different reason to jump headlong into 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. Perhaps at first, they completely missed the mark, because you'd preemptively sliced off that part of the space of possibilities, like my anecdote with Ladislav Špaček? Ironically, this is especially likely the more of a background one has in semantics and related disciplines, because such background typically causes one to approach new information with a more constrained set of possibilities. A neat example of this is a footnote where Ramscar responds to reviewers. The context is this:

> As I noted at the outset, after half a century of motivated effort, researchers have singularly failed to come up anything approximating a half-coherent empirical account of what a 'unit of meaning' is supposed to be (Ramscar & Port, 2015), and philosophical analyses that have long suggested that 'meaning units' are a fundamentally misguided idea (Wittgenstein, 1953; Quine, 1960; see also Fodor, 1998). Which is to say that although most linguists (and other researchers in the brain and cognitive sciences) clearly believe in compositionality,

> theoretical accounts of compositional meaning themselves offer nothing beyond blind faith, vagueness, and / or mysticism. (bibliography references in the original, Ramscar 2019: 57–8)

To which Ramscar adds the following footnote:

> A previous version of this work was criticized for making "little connection with relevant current work in theoretical pragmatics" and "not much connection with the state of the art in theoretical linguistics, pragmatics, psycholinguistic processing, or children's semantic/pragmatic development," while a reviewer complained that "the specific cases ... discussed are all to do with words (names, verbs, nouns, gender systems), while syntax, the key driver of linguistic compositionality, is not mentioned." What I hope is clear to the careful reader (and even future reviewers) is that I hold out no hope that a successful theory of human communication can be built on the idea of 'units of meaning' at any possible level of description, and that as a result, I have little to say about work founded on this idea (the 'state of the art in theoretical linguistics,' 'pragmatics,' 'syntax') other than to note that if the foundations of a scientific theory are wrong, it seems reasonable to assume that its ultimate contributions to human understanding are likely to be minimal. (Ramscar 2019: n. 14)

Let's put aside now whether some of these disciplines, particularly pragmatics, may not harbor some surprising allies – possibly not with the same level of theoretical clarity and sophistication, but at least people aware that being too literal about locating meanings in words rather than minds is an enterprise fraught with danger. I'd rather not speak to the current state of pragmatics, as it's not my area

of expertise, but we have seen that some historical figures who have influenced the roots of the field were headed in roughly the same direction. Regardless, the quote oozes frustration and exasperation, and depending on your perspective, may even sound somewhat presumptuous, but the basic sentiment should feel relatable by now: "I really made an effort and gave you thousands of words to latch onto, instead of dropping a few enigmatic bits of wisdom and calling it a day; I know that's no guarantee of getting my meaning across, indeed, this is one of the key points I'm trying to make in my account of how meaning works; in the future, I'll try and come up with other ways to approach the subject, and perhaps the nth attempt will be the charm; but at this point, I'm just tired and annoyed I'm not getting through."

In less lyrical terms: the more we know about a subject – the better we're able to predict what others might be trying to tell us about it *within* its confines – the more likely we are to be caught off-guard when called upon to internalize a related but truly novel concept.[22] Metaphorically speaking, in such situations, the lump of stone we start with tends to already have the key target bits sliced off, and having listened or read for a while, busying ourselves away at the chisel, we end up bewildered and empty-handed. We then have to undergo the laborious process of retracing our steps, throwing away our painstakingly acquired assumptions and preconceptions, and starting again from a much broader picture, slowly chipping away at a much larger lump. Even if you're not fully convinced yet, even if you're still harboring doubts and reservations and pondering counter-arguments, then if what I'm saying started out feeling like gibberish, but is now sort of making sense,

[22] This is an extension to the point made by Yudkowsky (2015: 37–9) and cited earlier: trying to explain something to someone who has little of the required background knowledge may be an uphill struggle, but addressing people who *think* they *do* have all the background knowledge necessary is even worse.

at least in places – then deep down, you already know it's true.

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.

We sometimes fancy ourselves as impartial, objective observers, but the fact that meaning is never "out there" but always "in us" should be a sobering thought. This is not to say that everything is relative and intersubjective agreement is impossible. To the contrary, it's very much possible and happens on a daily basis, and language is the superpower that helps us achieve it. But again, monologue is about the hardest way to achieve it. And any linguistic methodology which analyzes speakers's behavior in communication from an outside vantage point – be it as well-intentioned and rigorous as it can possibly get – is still fundamentally a one-sided affair. Will this ultimately lead us to bring new methods to the forefront? In conversation analysis, analytic insights are somehow triangulated via follow-up debriefing interviews with participants.[23] What would be a more general equivalent, when access to specific participants is impossible or impractical? Crowd-sourcing a distribution of opinions on the material at hand, sampled as independently as possible, rather than a point

[23]Note the plural there – participants. No one has the full context, not even each participant individually (hence misunderstandings and fights). Much less a linguist external to the situation, of course.

estimate by a single person or group of highly correlated people (the team behind a given piece of research)? This is also already being done, but in the context of a discriminative approach to meaning, these methods should be promoted a few ranks in terms of epistemic importance: suddenly, it's not a case of asking people on their subjective opinions about the meanings of words, it's asking for the real thing.

To reiterate what I said at the start: dictionaries can certainly be useful tools. Even compositionality can, for a given task – useful models typically have simplifying assumptions, it's about picking acceptable ones given the task. You just need to be aware that they're not the real thing, much like phonemes aren't. They're a map – one possible conceptualization of reality, with both advantages and drawbacks – not the actual territory (cf. e.g. Yudkowsky 2015: chap. 35 for a discussion of this distinction). So if you run into apparent paradoxes in trying to account for how meaning works, you should remember you can always throw away these particular maps, and use more appropriate ones.

I mentioned at the beginning that Gary Marcus is still clinging to the map which says, broadly, no human-level artificial intelligence without compositionality. To close with an interesting counterpoint, I offer the perspective of someone writing under the pseudonym nostalgebraist, who read and appreciated Marcus's book when it first came out in the early 2000s, but got a different takeaway from tinkering with the last decade's worth of fascinating developments in deep learning:

> GPT-2 can *fucking write*. (BTW, since we've touched on the topic of linguistic nuance, I claim the expletive is crucial to my meaning: it's one thing to merely put some rule-compliant words down on a page and another to *fucking write*, if you get my drift, and GPT-2 does both.)

> This should count as a large quantity of evidence in favor of the claim that, ***whatever*** necessary conditions there are for the ability to ***fucking write***, they are ***in fact satisfied*** by GPT-2's architecture. If compositionality is necessary, then this sort of "deep learning" implements compositionality, even if this fact is not superficially obvious from its structure. (The last clause should go without saying to a reader of ***The Algebraic Mind***, but apparently needs explicit spelling out in 2019.)
>
> On the other hand, if "deep learning" cannot do compositionality, then compositionality is not necessary to ***fucking write***. Now, perhaps that just means you can run without walking. Perhaps GPT-2 is a bizarre blind alley passing through an extremely virtuosic kind of simulated competence that will, despite appearances, never quite lead into real competence.
>
> But even this would be an important discovery – the discovery that huge swaths of what we consider most essential about language can be done "non-linguistically." (nostalgebraist 2019)

Being a very cautious, look-before-you-leap type of person by nature, I fully sympathize with the circumspection and hedging. That being said, this is exactly the point where you throw away the old map.

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 et al. 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).

The data gathering methodology follows the practice established with earlier spoken corpora built at the Czech National Corpus, starting with ORAL2006 (Kopřivová & Waclawičová 2006). Conversations were recorded on a portable device by one of the participants themselves, so without direct intervention from a linguistic expert or other outsider who might make the speakers uneasy. If possible, participants other than the one setting up the device had no prior knowledge of being recorded; they were instead informed post hoc, at which point their consent to keep and use the recording was of course sought.

This was done so as to minimize any effects the knowledge of being recorded

might have on the speech they produce. Even so, there were cases in which this preferred timeline of events was impossible to apply, but the general consensus in these matters seems to be that potential major adverse effects stemming from the observer's paradox should evaporate in a matter of minutes from the start of the recording, as participants shift their focus to the conversation at hand instead of its circumstances. This means that length is an additional safeguard against these effects. In the ORTOFON project, recording length was soft-limited to around 40 minutes, both to ensure diversity and to have manageable units of work for the manual stages of data processing. The resulting interquartile range of recording lengths in the ORTOFON v2 corpus is roughly from 13 to 26 minutes, with a median of 20, i.e. the length in minutes of the bulk of the data is in the double digits.

From a technical point of view, unlike in the case of Spoken BNC or Spoken BNC2014, participants were provided with a standardized recording device instead of being told to use their own (e.g. smartphones), in order to make sure recordings are comparable in terms of technical quality. While the details of the environment can hardly be controlled for in this type of material (e.g. placement of the recording device, relative position of speakers, sources of background noise etc.), it is still helpful for further processing if the data are as homogeneous as possible. In this case, it meant using the same type of microphone across all recordings and the same format (uncompressed WAV files in LPCM format, sampled at 44.1 kHz with a depth of 16 bits). The first criterion is impossible to achieve when letting all participants use a device of their own choosing, and the second one is non-trivial, seeing as built-in recording apps on smartphones tend to favor simplicity in their user interfaces and often use a compressed audio format without the option to switch.

The recordings were then manually transcribed using ELAN (Sloetjes & Wittenburg 2008) on two main tiers, a basic transcript and a matching phonetic layer, split into time-aligned segments of 25 words at most. Additional annotation includes paralinguistic and ambient sounds. A short stretch of transcript as it appears while editing within ELAN is given in Figure 4.1 as an example.

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

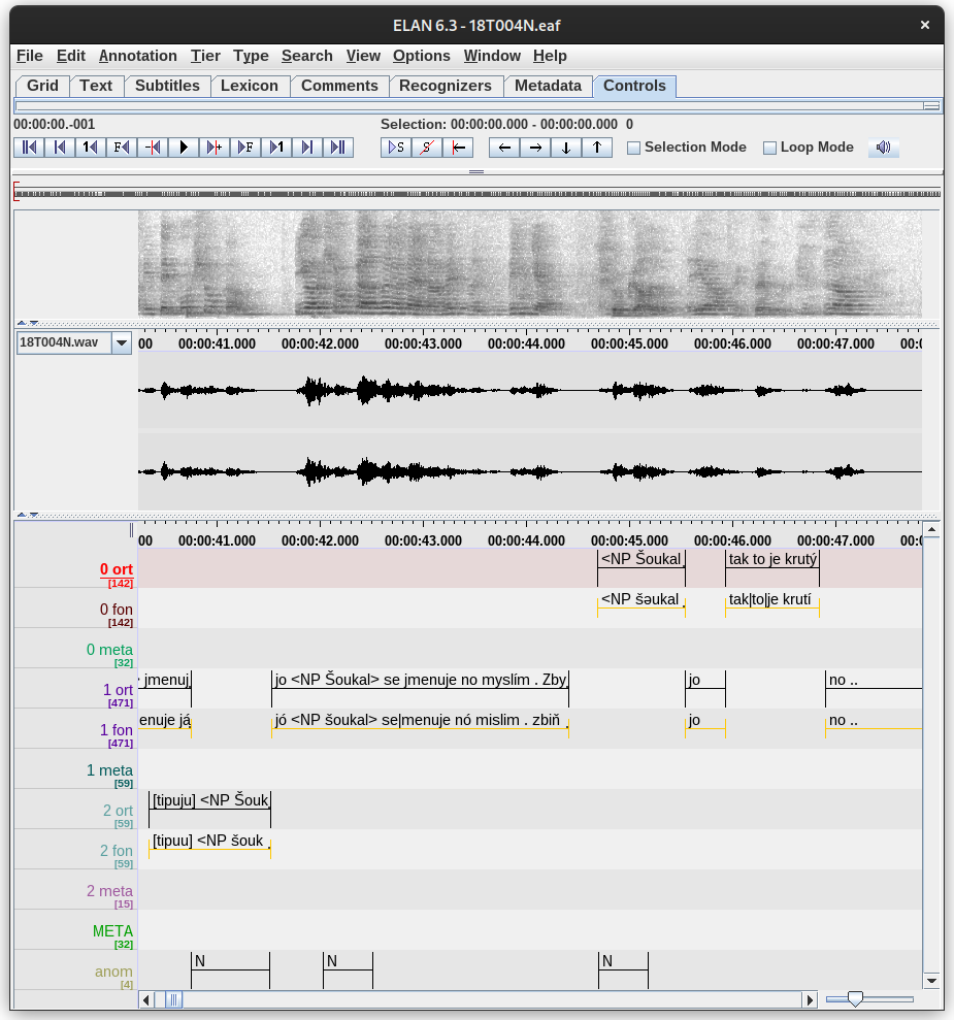

Figure 4.1: Sample transcript of recording 18T004N from the ORTOFON corpus, as opened in the ELAN transcription program. Transcription tiers prefixed with numbers are speaker-specific, each speaker gets his or her own copy of that tier: an `ort` tier with the basic transcript, a `fon` tier with the corresponding phonetic transcript, and a `meta` tier with paralinguistic annotation. The two last tiers are shared: `META` (situation-related non-speaker-specific events) and `anom` (anonymization; these stretches will be replaced with beeps prior to release).

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.

It would be a shame to completely give up on applying Prosogram, just because some of the input data has bad audio quality and the output will be rubbish in those cases. As shown in Section 2.2, the output can still be genuinely useful in many others. But it does mean we need to proceed with some caution.

As for data availability, all CNC corpora are generally made available through the online search interface KonText[1] as soon as ready for public use. This does however come with some restrictions compared to downloading the data locally. While providing such downloads for written corpora is generally problematic due to licensing restrictions, with spoken corpora, this is typically possible, provided that the data is properly anonymized, and we strive to provide download options when feasible, either on a per-request basis, or even in public data repositories such as LINDAT/CLARIAH-CZ.[2] In the case of the ORTOFON and ORATOR corpora, only ORTOFON v1 is currently publicly available for download via LINDAT/CLARIAH-CZ (Kopřivová et al. 2017a; Kopřivová et al. 2017b). In the future, we plan to increase our offerings in this domain; in the meantime, data from ORATOR or newer versions of ORTOFON can be requested privately via the CNC CLARIN K Centre helpdesk.[3]

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

[1] See https://korpus.cz/kontext.
[2] See https://lindat.mff.cuni.cz.
[3] See https://korpus.cz/clarin/helpdesk.

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, 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[4], 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[5] (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 [e͡ɪ] 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

[4] See e.g. `https://clarin.phonetik.uni-muenchen.de/BASWebServices` (Kisler et al. 2016) or `https://mowa.clarin-pl.eu`.

[5] See `https://montreal-forced-aligner.rtfd.io`.

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 [pi̥ɛ].

Now, ASR toolkits are generally able to accommodate multiple pronunciations per word form (see Figure 4.2 for a rough overview of how this fits among the other components involved in an ASR system), 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 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

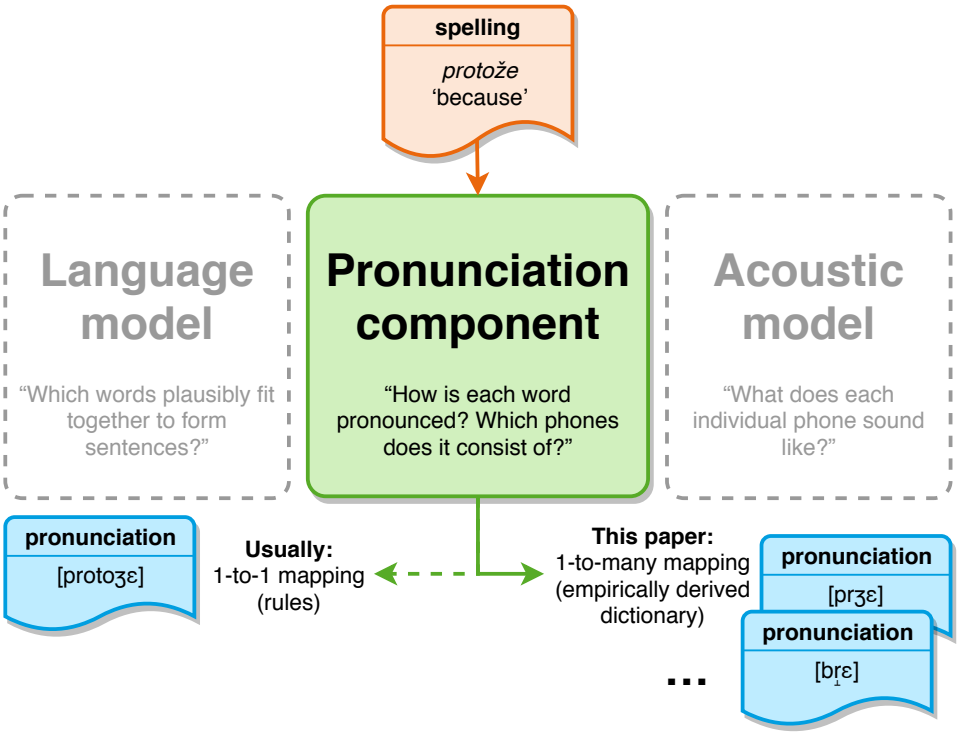

Figure 4.2: A simplified overview of the components of an ASR system, focusing on how pronunciation is handled. Originally prepared for Lukeš et al. (2018). "This paper" refers to that original place of appearance; disregard it.

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_protoʒɛ` or `protože_bʒɛ`, 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_protoʒɛ`, another one for `protože_bʒɛ`, 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[6] (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.

A caveat is in order here: unlike ORTOFON, the ORATOR data therefore contains no formally reduced pronunciation variants, not where appropriate, nor anywhere else. This can bias certain types of comparisons between the two corpora – basically any comparison that relies on information associated with specific phones. This includes phone durations (e.g. when measuring articulation rate), or pitch targets or Polytonia tones associated with specific nuclei. If required by the dictionary, the forced aligner will happily cram in all of the phones of the full pronunciation, even though a human transcriber would clearly identify this specific occurrence as reduced. In the results presented below, I'll be therefore avoiding such analyses, although they are definitely possible, either by restricting oneself to

[6] See `https://corpy.rtfd.io/en/stable/guides/phonetics_cs.html`.

each corpus individually, or by devising mechanisms to compensate for this type of bias.

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.[1] 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 5.1 and 5.2.

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.

[1] For a full overview, see Section 5.2 of Prosogram's User's Guide (Mertens 2020).

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.

First of all, some data got shaved off just by applying Prosogram, because Prosogram ran into errors or yielded no output. This was the case for about 10% of dialogue spans, which seems a lot, especially compared to monologue. However, taking a look at the corresponding number of words, it's only about 2%, which seems to indicate these were mostly short spans, probably containing backchannels such as *hmm*, as is typical of conversational speech.

Next, subjective estimates of audio quality are actually available for both corpora, on a scale from 1 (best) to 3 (worst). These are unfortunately not very reliably filled out, but a basic rule of thumb is that while 1 does not guarantee especially good quality, 2 and 3 *do* indicate pretty bad quality, so I excluded those, with rates relatively similar across both corpora. This audio quality estimate is also a global indicator, so it doesn't help us to pick out potentially problematic spans within a recording. For that, we can use the presence of unclear words or speaker overlaps as proxy local indicators of (worse) audio quality. Unclear words increase the likelihood of an inappropriate phone alignment, especially if the transcriber completely gave up on taking a guess at what they might be and instead provided just a rough estimate of their number, and overlaps are problematic both for forced alignment and F0 detection for obvious reasons. As Table 5.2 shows, the removal of spans with overlaps is the cleanup stage with the most drastic reduction of the size of the data, almost 40% of spans/words in dialogue, as they are quite common in conversational speech. By contrast, in monologues, they hardly occur at all, unsurprisingly so.

Moving on, I also removed spans with low average intensity. This cleanup step had more of an impact in previous iterations of the analysis, but since it ultimately

Table 5.1: 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 5.2: Same as Table 5.1, but relative numbers, giving proportions shaved off by individual cleanup steps. Also, unlike in Table 5.1, 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

led me to implement span pre-amplification via normalization, its effect has become relatively marginal. Still, I remove the lowest 1st percentile of spans, or those with a mean intensity peak per nucleus of less than 70.4 dB.

As a technical aside: Praat and therefore also Prosogram claims to provide intensity measurements in dB SPL (sound pressure level), but this is somewhat confusing in the context of recorded audio, as there is no intrinsic loudness in the signal: the loudness depends purely on how loud you play it back, and this can vary with each playback. In digital audio, the customary practice is to talk about dBFS (full scale), a measure of intensity relative to the maximum intensity that can be represented at a given bit depth. In order to properly map between dBFS and dB SPL, a calibration sound must be recorded ***in the same recording conditions*** and its SPL measured with a sound level meter. Praat requires no such calibration sample and instead hardcodes the conversion in a way that makes the outcome fall in a roughly "reasonable" range for most speech recordings.[2] Obviously, this heuristic fails when the speech is faint because recorded from afar, not because spoken softly; hence, pre-amplification.

An artifact that often plagues F0 detection is so-called octave jumps. These happen when the F0 detection algorithm (typically auto-correlation) has trouble latching onto the correct frequency, either because the periodicity of vocal fold vibration is extremely low or irregular (as happens with creaky voice phonation), or when background noise or a higher frequency component of speech is mistaken for F0 because the auto-correlation is stronger. This results in a zigzaggy detected F0 curve which jumps back and forth between high and low values in a way that's unrepresentative of the corresponding auditory perception. What partly helps with this is defining a stricter frequency range for F0 detection, as I did above, which

[2] Personal communication by Tomáš Bořil, Institute of Phonetics, Faculty of Arts, Charles University.

prevents suspiciously high (or low, for that matter) values from being considered as viable candidates, and also the fact that Prosogram only looks at F0 values within what has already been identified as a syllabic nucleus (unlike Momel), but in general, Prosogram isn't immune to octave jumps. In fact, one instance can even be seen in Figure 2.1: the lone thick black stroke outside the prosogram's axes on the initial vowel in *oslavit*. It seems they could at least partially be ironed out during the stylization, but this would in turn require some heuristics as to which kinds of F0 variation to flag as spurious, so I understand the decision to take the F0 curve, once extracted, at face value.

However, this doesn't mean we can't implement such heuristics on our own. One possible strategy to identify spans with suspiciously zigzaggy F0 patterns is by calculating the so-called normalized pairwise variability index (nPVI) (Grabe & Low 2002: 520), of the discretized F0 contour:

$$nPVI = \frac{100}{m-1} \times \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right|$$

Where:

- m is the number of elements in the sequence (pitch measurement points, in this case)
- d_k is the value (pitch, in this case) measured for the k^{th} element of the sequence

Roughly speaking, the nPVI is a measure of cumulative local variability. For instance, let's imagine two F0 curves, both confined in the range between 200 and 150 Hz. One just gradually declines from the maximum to the minimum, the other keeps jumping back and forth between them. Both will have the same pitch range, which is a measure of global variability, but the latter one will have a much higher nPVI, because the average change between two neighboring nuclei will be much

more pronounced.

To borrow an example from myself (Lukeš 2014: 36–7), an intuitive grasp of nPVI can be gained via a comparison with another measure of variation, the coefficient of variation (CV), i.e. the ratio of the standard deviation (s) to the mean ($\bar{x}$):

$$CV = \frac{s}{\bar{x}}$$

First of all, both measures are relative, so they can be used to compare sequences of varying lengths. It would be no use if e.g. nPVI grew with the length of the sequence, as it would make then be more likely to exclude longer spans. Second of all, Figure 5.1 shows a toy example comparing two sequences with the same mean (5), but different nPVIs and CVs. In sequence 1, the values are much more spread out, resulting in a higher standard deviation, and consequently, a higher CV than in sequence 2 (0.404 for sequence 1 vs. 0.211 for sequence 2). Conversely, the values in sequence 2, while keeping close to the mean, oscillate wildly back and forth, resulting in a higher nPVI (0.400 for sequence 2 vs. 0.154 for sequence 1). In other words, nPVI and CV are sensitive to different kinds of variability: CV is **global**, it quantifies the overall range covered by the sequence, whereas nPVI picks up on **local** fluctuations. As far as F0 contours are concerned, global variation is fine, over enough words, speakers can cover quite a lot of intonational ground. But they don't typically jump frantically up and down from nucleus to nucleus: local fluctuations are thus suspicious and indicative of possible octave jumps. Hence the use of nPVI to detect them.

The kernel density plot of F0 nPVI per span for the combined data set (ORTO-FON and ORATOR together) is given in 5.2. Based on cursory inspection, I decided to remove the upper 5th percentile of spans with the highest nPVIs, i.e. spans with

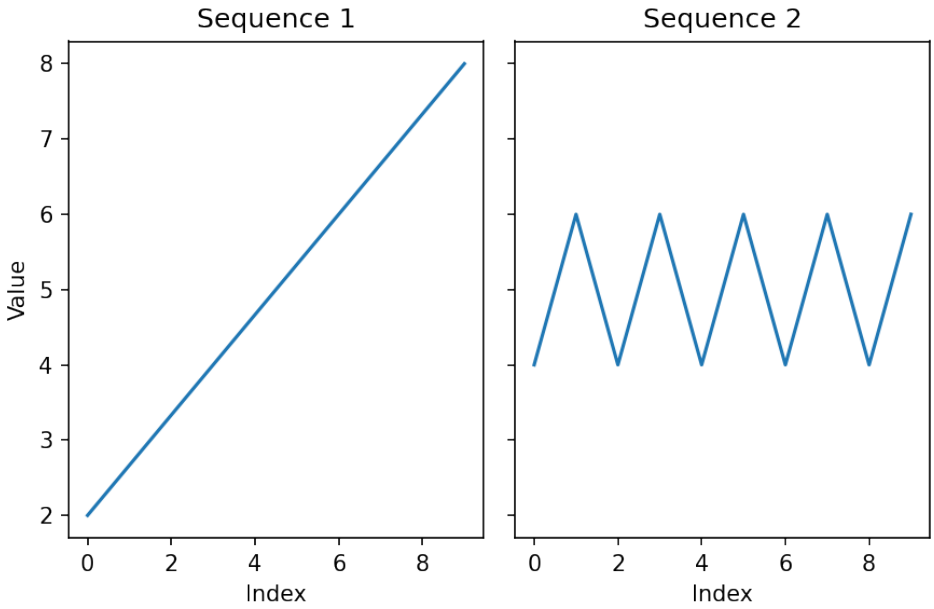

Figure 5.1: A toy example showing two sequences with the same mean (5), but different nPVIs (0.154 vs. 0.400) and CVs (0.404 vs. 0.211).

nPVI approximately 0.32 and higher. Two examples of actual F0 curves from the data set, are given in Figure 5.3. According to the ultimately selected threshold value for acceptable nPVI, the right one is borderline acceptable, while the left one will be clearly rejected. Manual inspection of the corresponding sound samples confirms this is a good thing: in reality, the signal contains no such wild F0 fluctuation.

Another possible heuristic is to throw away recordings made under conditions which are known to cause problems with F0 detection. Based on available metadata, I identified two such conditions: car rides and Skype calls. Car rides tend to be extremely noisy, including sources of continuous, unbroken noise like the constant underlying hum of the engine or of tires rolling on the road. As for Skype, the compression for real-time Voice over Internet Protocol (VoIP) audio can be drastic, as it primarily cares about acceptable intelligibility and low latency, not a faithful representation of the input frequency spectrum. Consequently, the results of auto-

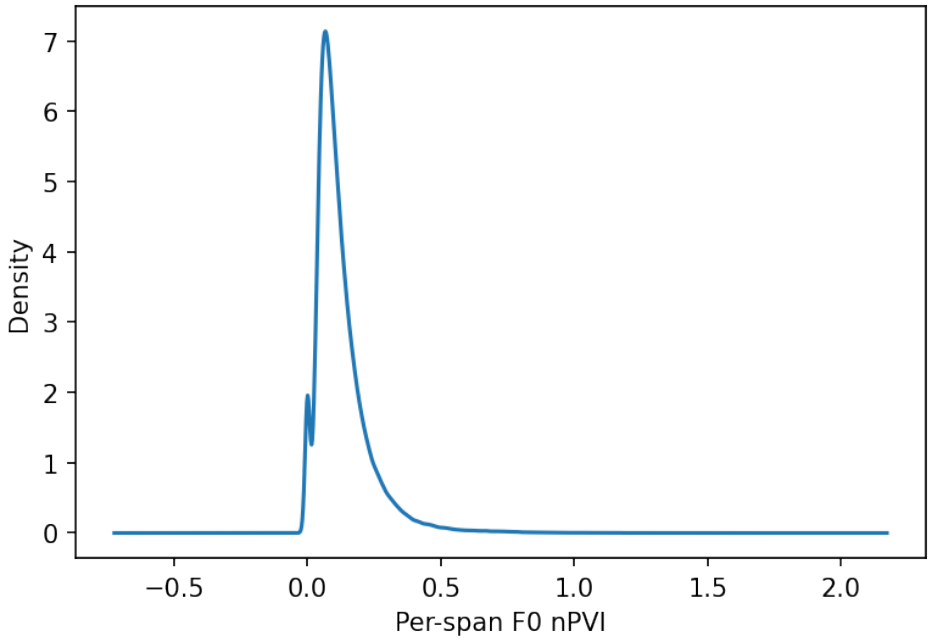

Figure 5.2: Distribution of per-span F0 nPVIs in the combined data set (both ORATOR and ORTOFON).

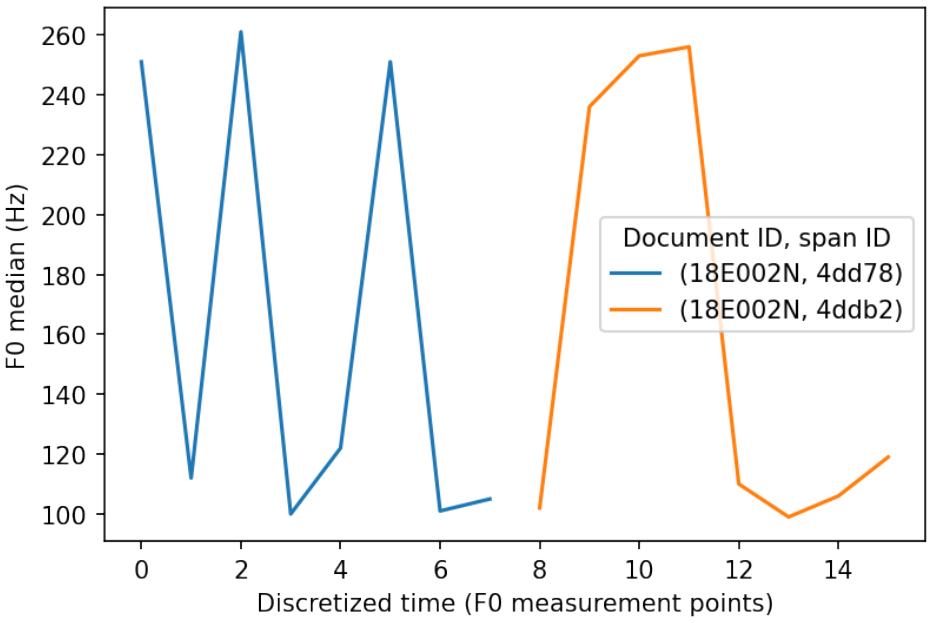

Figure 5.3: Two F0 contours from the ORTOFON corpus, with respective nPVIs of 0.61 and 0.28.

matic F0 extraction tend to be problematic in both cases. These conditions only affected a small minority of recordings under the dialogue setting, but out they all went.

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.

The cleanup steps detailed above were all relatively low-hanging fruit. The accuracy of the cleanup could possibly be further improved via some sort of automated sound quality labeler, whether based on manual heuristics or machine learning. This would even allow a rolling window quality measure, allowing to assign different grades to different portions of a recording and perform a more granular selection. Such a system is however a research project in its own right.

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, Huber & Sussman (2011), Figure 1.

In our data, we unfortunately don't have age information about the monologue speakers. Therefore, Figure 5.4 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.

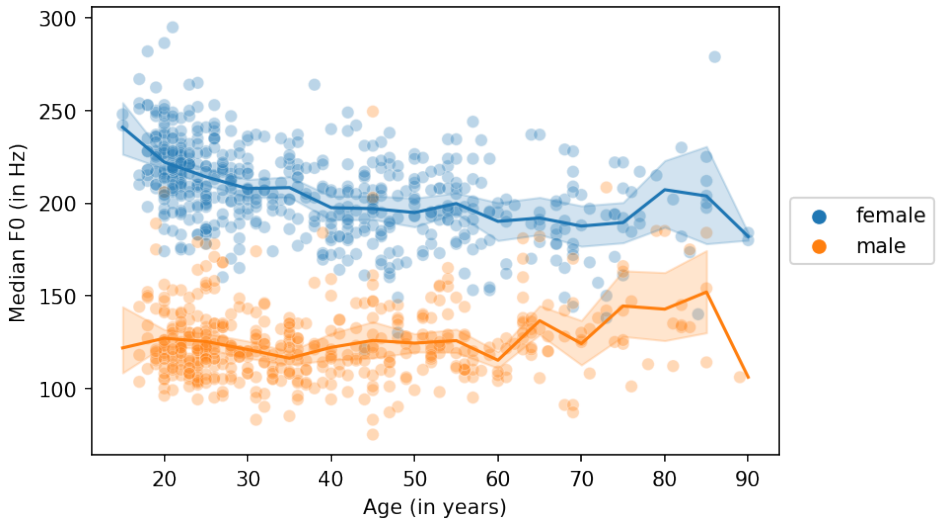

Figure 5.4: Median F0 per speaker in recording, in Hz, broken down by gender and age. Only covers dialogue speakers.

Not to exclude the monologue data, Figure 5.5 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.

Another possible reason is that the shift is somehow due to differences in composition of the two subcorpora in terms of speakers' age. As mentioned above, the monologue speakers' ages are not known, but a very rough estimate would be that there will be a lower proportion of younger people, purely because of situational context: we've previously established that the monologues are mostly lectures, which will bias the speakers towards middle age. By contrast, about a third of the dialogue data is by under-25-year-olds. This means differences in age composition could particularly affect the results for women, for whom Figure 5.4 suggests a decrease of median F0 between their twenties and later stages in life. However, young women with high F0 push the distribution rightward, thereby acting to close the gap we empirically observe. Given that the distribution for women in monologues is even further rightward and the gap is still there, in spite of the speakers being presumably older on average, I find age an unlikely explanation for the gap.

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

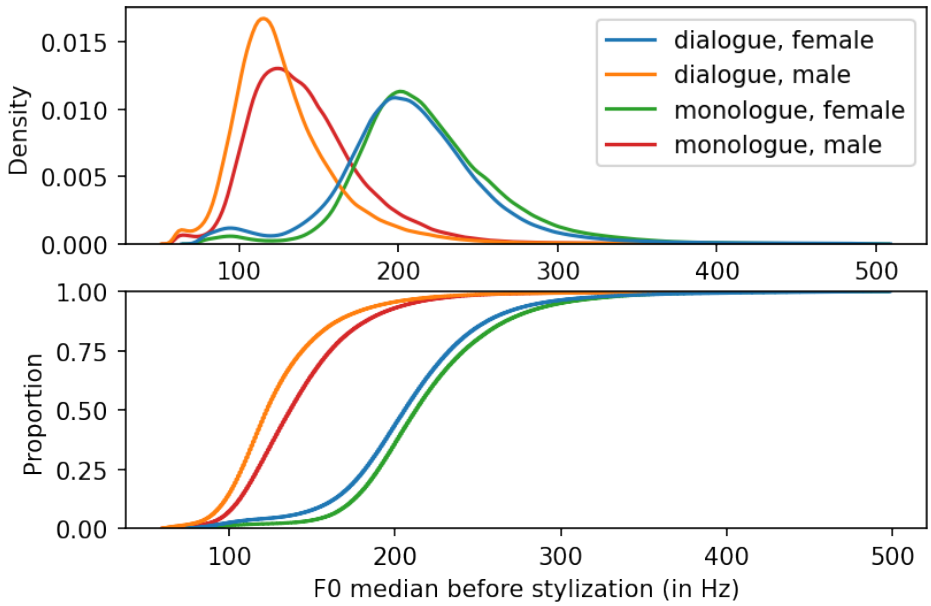

Figure 5.5: Distributions of median F0 per nucleus before stylization: kernel density estimate (top), empirical cumulative distribution function (bottom).

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 5.6; 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.

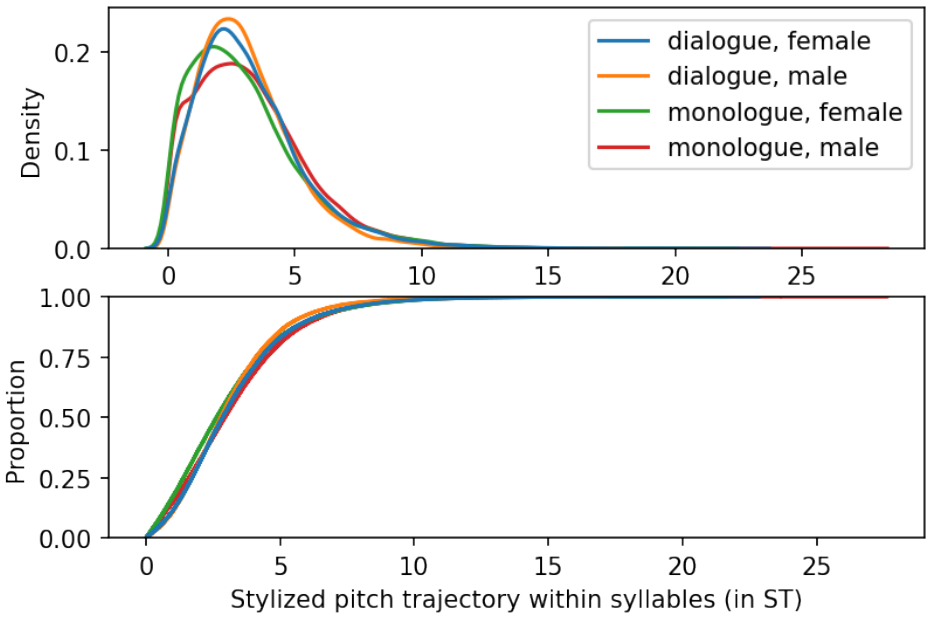

Figure 5.6: Distributions of glissandos (cumulative pitch trajectory per syllable) in both dialogue and monologue data: kernel density estimate (top), empirical cumulative distribution function (bottom).

An interesting observation results from looking at the *proportion* of glissandos, i.e. syllable with `trajectory` greater than 0, as shown in Table 5.3. 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.

Table 5.3: 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

This is but a brief look at glissandos in the material, but it warrants digging deeper in the future. Possible areas to explore include distinguishing between different types of glissandos, based on prevailing direction (up vs. down) or amplitude, and unpacking the relatively broadly defined categories (monologue vs. dialogue, female vs. male) using more fine-grained document- and speaker-level metadata.

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*[3] 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 5.1 presents the results of fitting an ordinary least squares regression model to the entire data set, with KIND and GENDER as predictors. Figure 5.7 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

[3] See https://www.statsmodels.org/.

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.

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

For easier orientation, Table 5.4 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 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

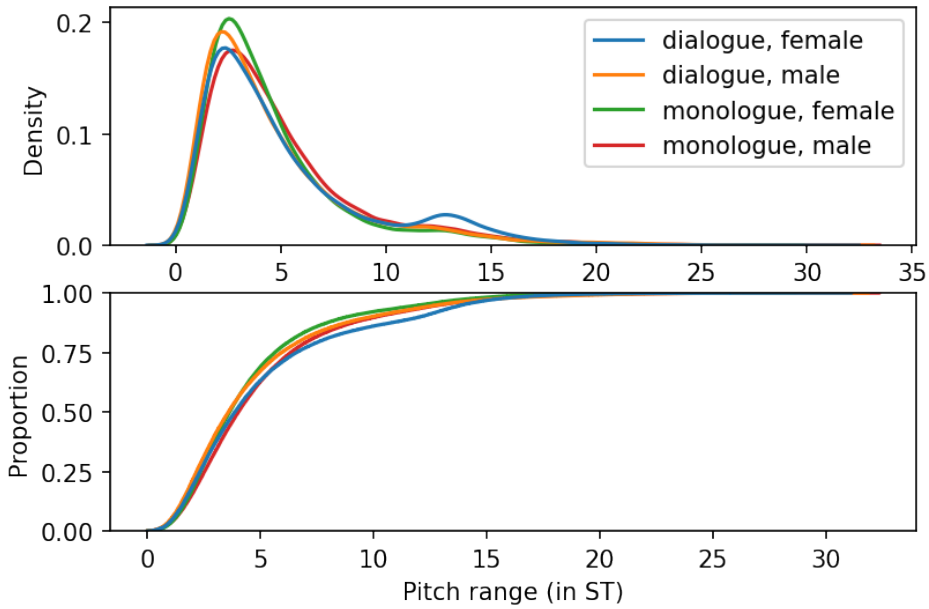

Figure 5.7: Distributions of interpausal pitch ranges in both dialogue and monologue data: kernel density estimate (top), empirical cumulative distribution function (bottom).

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 5.4: 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

Let's move on to the model focusing on monologues only. This time, we have a mixed effects model, with GENDER, INTENDED AUDIENCE and FRAMING as fixed effects, and SPEAKER as a random effect. INTENDED AUDIENCE can take on one of two values: public and restricted. This aims to distinguish between events that were open to any member of the public, at least in principle, and those that were meant for a specific, restricted group. FRAMING is somewhat vague as umbrella terms go, but the specific values will hopefully make it clearer what is meant by it: official, political, popular, professional and scientific. It constitutes another dimension along which the different contexts in which speeches are given can be distinguished. As there was way less data in the political FRAMING than in the other ones, I excluded it from the analysis. An overview of the results is given in Listing 5.2. Apart from the tendency to a wider pitch range in men, already noted in the previous model, there are no particularly convincing effects. INTENDED AUDIENCE comes relatively close, with a slightly wider pitch range when addressing a restricted audience, but the effect size is small and the 95% confidence interval includes 0, if just barely. No clear trends related to FRAMING emerge.

```
                Mixed Linear Model Regression Results
========================================================================
Model:                  MixedLM      Dependent Variable:      range
No. Observations:       153397       Method:                  REML
No. Groups:             424          Scale:                   11.2842
Min. group size:        1            Log-Likelihood:          -404288.8049
Max. group size:        2775         Converged:               Yes
Mean group size:        361.8
------------------------------------------------------------------------
                                Coef.  Std.Err.   z    P>|z| [0.025 0.975]
------------------------------------------------------------------------
Intercept                        4.475    0.152 29.437 0.000  4.177  4.773
gender[T.male]                   0.445    0.159  2.794 0.005  0.133  0.757
intended_audience[T.restricted]  0.145    0.083  1.740 0.082 -0.018  0.307
framing[T.popular]              -0.014    0.110 -0.126 0.900 -0.230  0.202
framing[T.professional]         -0.172    0.138 -1.242 0.214 -0.442  0.099
framing[T.scientific]            0.023    0.130  0.177 0.860 -0.232  0.278
Group Var                        2.121    0.047
========================================================================
```

Listing 5.2: Mixed effects model of `pitch range ~ gender + intended_audience + framing` in the monologue subset of the data.

Finally, let's 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 (see Figure 2.3 for a map overview). Results of the fit are summarized in Listing 5.3 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.[4] 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

[4] I don't know of any study of acoustic correlates of Czech stress focusing specifically on speakers of this particular dialect, but I would wager that it would find some, unlike the results reported by Skarnitzl (2018) and cited previously, which found none. At the very least, since there is no phonemic vowel length contrast, vowel duration is freed up to participate as a cue for stress contrasts.

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

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 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.

Another important point is that these analyses should be seen as exploratory, as a point of departure an inspiration for more targeted, better controlled studies. With a few exceptions, the set of speakers in the monologue and dialogue conditions doesn't overlap. To get better, more robust insights into how these conditions affect speaking styles, it would be a good idea to design a follow-up study with paired observations, i.e. a much more controlled corpus with both speaking styles produced by the same speakers under different conditions. So one goal of the present study is to encourage researchers to pull on these threads and provide some prior estimate on how worth their while each of them might be.

This is a point I want to insist upon: at times, researchers have felt the need to step up in defense of their ability to build their own corpora of spoken Czech, at their discretion, even though the CNC provides a selection of general-purpose corpora in this domain (see e.g. Chromý 2017). This is understandable: funding opportunities are tight, and if a project proposal budget includes funds for data gathering, a referee might object whether new data is really needed. As hinted at above, my answer to that question is clear: if the project proposal claims so, it very

likely is. While a lot of research mileage can be had from general-purpose spoken corpora on their own, they are also extremely useful for quick intuition checks and proof-of-concept studies which determine further directions to explore on specially collected data.

A general-purpose spoken corpus can't be everything to everyone: decisions have to be made about the data gathering methodology, recording requirements (chiefly quality, recording device, separation of sources), transcription and annotation, etc. These decisions are also directly tied to scale: a smaller corpus can in principle afford richer annotation, more manual checks and more stringent inclusion criteria than a large corpus (millions of words, hundreds of hours of recordings). Consequently, general-purpose spoken corpora aren't quite the right fit for many kinds of research on spoken language, due to specific requirements related to e.g. experimental design, sound quality (especially in the case of minute phonetic analyses, whether auditory or instrumental), or metadata, as Chromý (2017) points out.

A concrete example, to make it quite clear what the results presented here should *not* be taken as: in the domain of forensic phonetics, extreme care must be taken when preparing population statistics for a given voice parameter, often used as grounds for comparison in case work. If the median F0 in two samples of speech is comparable, then the likelihood of the samples belonging to the same speaker crucially hinges upon how common such F0 values generally are in the population. In other words, if both samples are e.g. noticeably higher pitched than usual for the given gender, that constitutes some evidence in favor of identification; if they're similar but close to the central tendency in the population, they could easily come from two different people. Therefore, when building an F0 database for forensic purposes, audio quality and F0 extraction accuracy should be paramount concerns. For this reason, special purpose data is typically collected (Skarnitzl & Vaňková

2017).

Again, this is not to say that such statistics extracted from a general-purpose corpus cannot be useful even for forensic phonetics – as preliminary analyses meant to guide future work. One possible inspiration from the results presented here can be as to the need (or lack thereof) to focus on collecting separate population statistics of F0 for various Czech dialects. But using them directly in situations which can influence whether an individual will be convicted or not, without any caveat, is problematic, to say the least.

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.

At any rate, analogous analyses for casual spoken English would not only put the stereotype of comparatively dull Czech intonation into clearer perspective, they would also be relevant to a larger community of linguists, given the status of English as a global language. But even if you don't particularly care about Czech, I hope this has been an enjoyable tour of what's currently possible in terms of exploring spoken language at corpus scale.

There are various factors that could have influenced the results in unexpected ways, and that would warrant further investigation, either with specially collected

data, or possibly more sophisticated analyses of the same data. Since one of the key variables which I looked at is gender within various other categories, a pronounced imbalance in the distribution of genders across those categories could skew the results. While I've taken care to discard categories which are generally poorly represented overall, especially in the case of the ORATOR corpus, where no balancing attempt was made, resulting in some quite tiny divisions within the data, the possibility of spurious results is still there. Possible strategies to minimize it even further include randomized sampling/bootstrapping techniques.

Furthermore, as noted previously all the way back in Section 2.3, the *pohraničí* regions (POHRANIČÍ ČESKÉ and POHRANIČÍ MORAVSKÉ) are quite problematic as geographic entities, as after all can be seen from the map in Figure 2.3. They mix together areas which are quite far apart from each other in topographical terms, sometimes even non-contiguous, separated by other intervening dialect regions. Redoing the analysis after redrawing those boundaries, at least splitting the *pohraničí* regions into sub-regions based on which true inner dialect region is closest, might yield interesting insights and a more appropriate picture for the regional perspective on the data.

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.[5]

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

[5] This should not be taken as a criticism of Prosogram's reliability in comparison with MFA. My point is simply that the longer the dependency chain, the more points of failure it introduces. MFA by itself is by definition a simpler and therefore more reliable annotation system than MFA + Prosogram.

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.

Rather, the practical alternative is to sacrifice some of the detail and aggregate the information at the word level. This requires careful consideration of the tradeoffs, so that users are empowered to the degree that they can be, not less, but also so that they are not overwhelmed. One possibility that I'm currently exploring is concatenating the tones returned by Polytonia per word, and providing a word-level tonal transcript; another one is based on the idea of providing word-level information about pitch and duration which relates the characteristics of a given token to the distribution of those characteristics among all tokens of that type, by means of percentiles. For instance, users could search for instances of ***protože*** which have an unusually long duration (say in the upper decile) or wide pitch range.

That actually implementing prosodic annotation in a corpus search system is far from a straightforward issue is also shown by the challenges faced by (Pęzik 2018), who implemented a custom system where some data is actually computed on-the-fly. This circumvents some of the restrictions: one can use a word-level indexing engine, while "decorating" the results with sub-word-level analyses. However, it also makes it impossible to use for searches, or at the very least, makes such queries tricky and palpably slower. Personally, this is not an acceptable restriction, as searching is one of the most useful applications of prosodic annotation that I anticipate, based on personal experience in using it to explore the raw data. Furthermore a fully custom system is out of the question in the CNC context, as we strive hard not to force

users to learn new interfaces for different types of corpora. So prosodic annotation will have to fit into Manatee somehow.

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,[6] 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.

I mentioned previously that Daneš (1957) has many shrewd and accurate insights about spoken language. Since I have been quite critical of some key aspects of his approach, I would be remiss not to expand also on some of the elements I consider very apt and to the point, at least briefly. In many ways, Daneš actually succeeds in discerning the distinguishing features of the spoken mode, especially in comparison with earlier attempts. For instance, he correctly observes that the phrasing of speech into phonetic words and intonation units is not really bound by semantic restrictions, but rather by physiological criteria and rhythmical considerations

[6] To be fair, as we've seen, automatic prosodic annotation also has a variety of reliability issues, though of a different kind.

(Daneš 1957: 18, p. 23). This is in stark contrast to figures of the earlier generation like František Trávníček who insisted that the division of speech into units followed semantic boundaries, which is a typical insight to have if you only ***think*** about speech, but don't actually ***listen*** to any.

But perhaps the most praise that I can heap upon Daneš is that he keenly realized the dangers of the dictionary trap for linguists analyzing intonation, clearly stating that one should not expect a one-to-one mapping between functions, meanings or emotions on the one hand, and intonation patterns on the other (Daneš 1957: 159). The relative vagueness of intonation compared to words is actually an advantage here: it nudges us towards realizing that meaning is in the mind, and language just helps us discriminate it.

6 Conclusion

Going in, my intuition was that monologues would have more pitch variation than dialogues. This was mostly based on a stereotype that I would describe as "castle tour guide intonation". Castles, manors and the like are a major tourist attraction in the Czech Republic, creating a demand for tour guides. These are not typically professional speakers, and they're paid for repeating the same memorized text over and over to different groups of visitors. The result is a very distinctive speaking style, which often overlays adventurous yet wildly mismatched pitch patterns over the pre-supplied text, in a vain attempt to enliven the presentation. Looking back, this was probably an unrealistic expectation: as is often the case with stereotypes, they're striking and memorable, but not necessarily representative. Or rather, they might be representative in their immediate domain of application, but we have a tendency to overgeneralize them, as I did here. There was no reason to expect a preponderance of castle tour guide intonation in a corpus containing mainly lectures.

Reality turned out to be more complicated, as it often does. When gender is taken into account, opposite tendencies emerge in dialogue vs. monologue in terms of pitch ranges, with men in monologue / women in dialogue having a similar pitch range, which is higher than men in dialogue / women in monologue. Regional factors also seem to play a role, as does age to an extent, at least for men. Glissando

usage patterns further complicate the picture: the proportion of glissandos is higher in monologues and similar across genders this time, although men seem to use slightly more of them, while being also slightly more pronounced.

This is not to say that instances of this intuitively characteristic castle tour guide intonation don't show up in the data: they do. An example is shown in Figure 6.1,[1] where each row corresponds roughly to an intonation unit. You can see how in both units, the speaker starts off towards the top of her pitch range, then slowly drifts downwards towards the mid-range. This is especially nicely visible in the second, lower half of the prosogram. While downtrends within intonation units are of course the rule rather than the exception, this particular pattern of slowly sliding towards a tentative mid-range should be immediately recognizable to anyone who's heard it before. What it associates for me is an attempt at enlivening the speech, without having to think too hard about its contents, and also without committing overly much: start high, aim slightly downwards, set to cruise. It also shows how recognizable and distinctive intonation patterns are not necessarily those where a lot of variation happens – in this case, the speaker actually mostly *constrains* herself to the upper half of her pitch range, if anything.

So once prosodic annotation is available, there are many ways to rummage through the haystack in search of particular needles. So many in fact that one can't possibly explore all the options and possibilities on one's own. None of us have the full picture on any of the languages we speak (Ramscar 2019: 4 discusses this at length and proceeds to show how languages are statistically shaped to accommodate for that fact). None of us can therefore intuitively conjure up all the patterns it might possibly be interesting to search for and explore. This is why making prosodic annotation accessible to regular users of CNC corpora of spoken

[1] Note that I didn't find this example based on metadata, or by pure luck: it turned up in the process of targeted searches for interesting examples by filtering based on prosodic features.

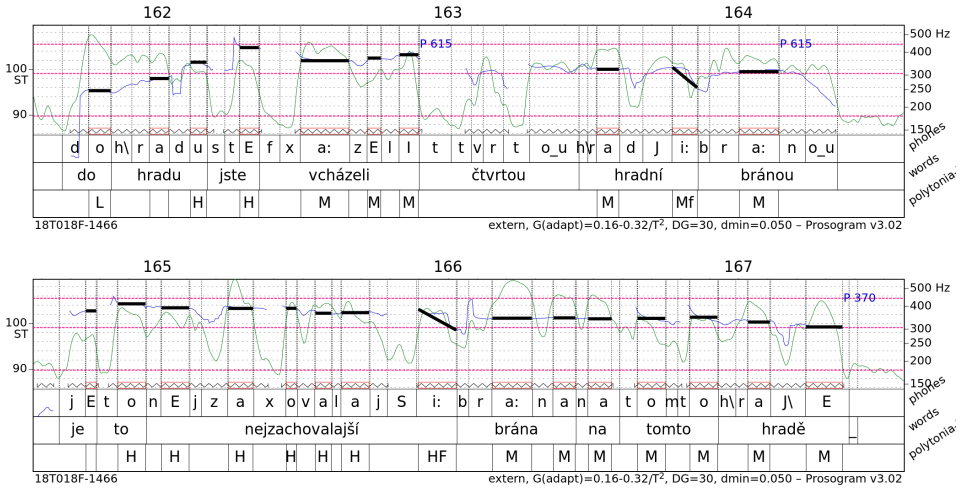

Figure 6.1: Prosogram of a female castle tour guide; each row is roughly a separate intonation unit. Notice how the speaker reaches the top of her intonation range early in each unit, and then gradually declines.

Czech is currently a priority. Czech linguistics has a great tradition of sophisticated, empirically driven analyses of spontaneous spoken language, see e.g. Müllerová (2022) or Čmejrková & Hoffmannová (2011), to mention just some of the key names in the field. Prosodic annotation in the spoken corpora of the CNC will hopefully enable researchers to build on this tradition and bring it even further.

Bibliography

Alexander, Scott. 2014. Beware Isolated Demands For Rigor. *Slate Star Codex*. https://slatestarcodex.com/2014/08/14/beware-isolated-demands-for-rigor/. (22 September, 2022).

Alexander, Scott. 2016. Superintelligence FAQ - LessWrong. *LessWrong*. https://www.lesswrong.com/posts/LTtNXM9shNM9AC2mp/superintelligence-faq. (22 September, 2022).

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

Arnold, Denis, Fabian Tomaschek, Konstantin Sering, Florence Lopez & R. Harald Baayen. 2017. Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE* 12(4). e0174623. https://doi.org/10.1371/journal.pone.0174623.

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

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *Mental Lexicon* 5(3). 436–461. https://doi.or

g/10.1075/ml.5.3.10baa.

Baayen, R. Harald, Yu-Ying Chuang & James P. Blevins. 2018. Inflectional morphology with linear mappings. *The Mental Lexicon*. http://www.sfs.uni-tuebingen.de/~hbaayen/publications/BaayenChuangBlevins.pdf. (19 August, 2018).

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins. 2018. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. https://doi.org/10.13140/rg.2.2.10821.55527.

Baayen, R. Harald, Peter Hendrix & Michael Ramscar. 2013. Sidestepping the Combinatorial Explosion: An Explanation of n-gram Frequency Effects Based on Naive Discriminative Learning. *Language and Speech* 56(3). 329–347. https://doi.org/10.1177/0023830913484896.

Baayen, R. Harald, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438–481. https://doi.org/10.1037/a0023851.

Baayen, R. Harald, Cyrus Shaoul, Jon Willits & Michael Ramscar. 2016. Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience* 31(1). 106–128. https://doi.org/10.1080/23273798.2015.1065336.

Baker, Gordon P. & P. M. S. Hacker. 1980. *Wittgenstein: Understanding and Meaning*. Oxford: Blackwell.

Balhar, Jan, Jarmila Bachmannová, Eva Balátová & others. 1999. *Český jazykový atlas* [*Czech Linguistic Atlas*]. Vol. 3. Praha: Academia.

Balhar, Jan, Jarmila Bachmannová, Eva Balátová & others. 2002. *Český jazykový atlas* [*Czech Linguistic Atlas*]. Vol. 4. Praha: Academia.

Balhar, Jan, Jarmila Bachmannová, Libuše Čižmárová & others. 2005. *Český jazykový atlas* [*Czech Linguistic Atlas*]. Vol. 5. Praha: Academia.

Balhar, Jan, Jarmila Bachmannová, Libuše Čižmárová & others. 2011. *Český jazykový atlas – Dodatky* [*Czech Linguistic Atlas—Addenda*]. Praha: Academia.

Balhar, Jan, Pavel Jančák, Jarmila Bachmannová & others. 1992. *Český jazykový atlas* [*Czech Linguistic Atlas*]. Vol. 1. Praha: Academia.

Balhar, Jan, Pavel Jančák, Jarmila Bachmannová, Libuše Čižmárová & others. 1997. *Český jazykový atlas* [*Czech Linguistic Atlas*]. Vol. 2. Praha: Academia.

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

Beckman, Mary E., Julia Hirschberg & Stefanie Shattuck-Hufnagel. 2005. The Original ToBI System and the Evolution of the ToBI Framework. In Sun-Ah Jun (ed.), *Prosodic Typology—The Phonology of Intonation and Phrasing*, 9–54. Oxford: OUP.

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.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language Models are Few-Shot Learners. arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Bréal, Michel. 1868. ***Les idées latentes du langage***. Paris: Hachette.

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

Bělič, Jaromír. 1972. ***Nástin české dialektologie***. Praha: Státní pedagogické nakladatelství.

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

Chambers, Jack K. & Peter Trudgill. 1998. ***Dialectology***. 2nd edn. CUP.

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.

Chromý, Jan. 2017. Comparison of spoken corpora from a sociolinguistic perspective. ***Slovo a slovesnost*** 78(2). 145–58.

Chuang, Yu-Ying & R. Harald Baayen. 2021. Discriminative Learning and the Lexicon: NDL and LDL. In ***Oxford Research Encyclopedia of Linguistics***. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.375.

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 Lingui-

stické). 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.

Goláňová, Hana. 2021. Delimitation of dialect regions, subgroups, areas and types in the Czech Republic. Charles University, Faculty of Arts, Institute of the Czech National Corpus. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4650. (6 September, 2022).

Goláňová, Hana & Martina Waclawičová. 2019. The Dialekt Corpus and Its Possibilities. *Journal of Linguistics/Jazykovedný casopis* 70(2). 336–344. https://doi.org/10.2478/jazcas-2019-0063.

Goláňová, Hana & Martina Waclawičová. 2021. Mapka: A map application for working with corpora of spoken Czech. *Journal of Linguistics/Jazykovedný casopis* 72(2). 502–509. https://doi.org/10.2478/jazcas-2021-0046.

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

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.

Karaś, Halina. 2010. Nowe dialekty mieszane. In Halina Karaś (ed.), *Dialekty i gwary polskie: Kompendium internetowe*. http://www.dialektologia.uw.edu.pl/index.php?l1=opis-dialektow&l2=nowe-dialekty-mie

szane.

Kisler, Thomas, Uwe D Reichel, Florian Schiel, Christoph Draxler, Bernhard Jackl & Nina Porner. 2016. BAS Speech Science Web Services – an Update on Current Developments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3880–3885. Portorož, Slovenia: European Language Resources Association (ELRA).

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 & Martina Waclawičová. 2006. ORAL2006: korpus neformální mluvené češtiny. Ústav Českého národního korpusu FF UK. http://www.korpus.cz.

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 Komrsková, David Lukeš, Petra Poukarová & Marie Škarpová. 2017a. ORTOFON v1: balanced corpus of informal spoken Czech with multi-tier transcription (transcriptions). http://hdl.handle.net/11234/1-2580.

Kopřivová, Marie, Zuzana Komrsková, David Lukeš, Petra Poukarová & Marie Škarpová. 2017b. ORTOFON v1: balanced corpus of informal spoken Czech

with multi-tier transcription (transcriptions & audio). http://hdl.handle.net/11234/1-2579.

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.

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.

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. 2014. *Percepční citlivost ve frekvenční a temporální doméně u hudebních a řečových stimulů*. Praha: Univerzita Karlova, Filozofická fakulta, Fonetický ústav MA.

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

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

Marcus, Gary. 2022. Did GoogleAI Just Snooker One of Silicon Valley's Sharpest Minds? Substack newsletter. *The Road to AI We Can Trust*. https://garymarcus.substack.com/p/did-googleai-just-snooker-one-of. (19 September, 2022).

Marcus, Gary F. 2001. *The algebraic mind: Integrating connectionism and cognitive science* (The Algebraic Mind: Integrating Connectionism and Cognitive Science). Cambridge, MA, US: The MIT Press.

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. 2022. *Karl Bühler's Organon model and the Prague Circle*. Mp3. https://hiphilangsci.net/2022/01/01/podcast-episode-21/. (14 September, 2022).

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). 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/.

Milin, Petar, Laurie Beth Feldman, Michael Ramscar, Peter Hendrix & R. Harald Baayen. 2017. Discrimination in lexical decision. ***PLOS ONE*** 12(2). https://doi.org/10.1371/journal.pone.0171935.

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

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. 1990. ***Change in Language: Whitney, Bréal, and Wegener***. Routledge. https://www.jstor.org/stable/415149?origin=crossref. (23 August, 2022).

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.

Nolan, Francis, Kirsty McDougall, Gea De Jong & Toby Hudson. 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. ***International Journal of Speech Language and the Law*** 16(1). 31–57. https://doi.org/10.1558/ijsll.v16i1.31.

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. Python.

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.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. http

s://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen & Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv. https://doi.org/10.48550/arXiv.2102.12092.

Ramscar, Michael. 2019. Source codes in human communication. *arXiv:1904.03991* [*cs, math*]. http://arxiv.org/abs/1904.03991. (17 May, 2020).

Ramscar, Michael. 2020. The empirical structure of word frequency distributions. *arXiv:2001.05292* [*cs*]. http://arxiv.org/abs/2001.05292. (19 January, 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.

Ramscar, Michael & Daniel Yarlett. 2007. Linguistic Self-Correction in the Absence of Feedback: A New Approach to the Logical Problem of Language Acquisition. *Cognitive Science* 31(6). 927–960.

Ramscar, Michael, Melody Dye & Joseph Klein. 2013. Children Value Informativity Over Logic in Word Learning. *Psychological Science* 24(6). 1017–1023. https://doi.org/10.1177/0956797612460691.

Ramscar, Michael, Melody Dye & Stewart M. McCauley. 2013. Error and expec-

tation in language learning: The curious absence of mouses in adult speech. *Language* 89(4). 760–793. https://doi.org/10.1353/lan.2013.0068.

Ramscar, Michael, Daniel Yarlett, Melody Dye, Katie Denny & Kirsten Thorpe. 2010. The Effects of Feature-Label-Order and Their Implications for Symbolic Learning. *COGNITIVE SCIENCE* 34(6). 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x.

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.

Rescorla, R. A. & A. R. Wagner. 1972. A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. F. Prokasy (eds.), *Classical Conditioning II: Current Research and Theory*, 64–99. New York: Appleton-Century-Crofts. http://www.columbia.edu/~rk566/Session4/Theory%20of%20Pavlovian%20Conditioning.pdf. (15 September, 2022).

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.

Sloetjes, H. & P. Wittenburg. 2008. Annotation by category—ELAN and ISO DCR. In ***Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)***, 816–820.

St. Clair, Michelle C., Padraic Monaghan & Michael Ramscar. 2009. Relationships between language structure and language learning: The suffixing preference and grammatical categorization. ***Cognitive Science*** 33(7). 1317–1329. https://doi.org/10.1111/j.1551-6709.2009.01065.x.

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(Historica-vol-29). 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.

Wells, J. C. 1997. SAMPA computer readable phonetic alphabet. In Dafydd Gibbon, Roger Moore & Richard Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems*, 684–732. Berlin, New York: Mouton de Gruyter. https://www.phon.ucl.ac.uk/home/sampa/.

Whitney, William Dwight. 1873a. Schleicher and the physical theory of language. In *Oriental and Linguistic Studies*, vol. I, 298–331. New York: Scribner, Armstrong and Co.

Whitney, William Dwight. 1873b. Steinthal and the psychological theory of language. In *Oriental and Linguistic Studies*, vol. I, 332–375. New York: Scribner, Armstrong and Co.

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.

Zaepernicková, Eliška & Martin Havlík. 2017. Prozodické aspekty reprodukované řeči v konverzačních příbězích. ***Studie z aplikované lingvistiky*** 8(2). 36–62.

Zipf, George Kingsley. 1949. ***Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology***. Cambridge, Massachusetts: Addison-Wesley.

nostalgebraist. 2019. human psycholinguists: a critical appraisal. Tumblr. ***Tumblr***. https://nostalgebraist.tumblr.com/post/189965935059/human-psycholinguists-a-critical-appraisal. (19 September, 2022).

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.

Glossary

ASR Automatic Speech Recognition

CLA Czech Linguistic Atlas

CNC Czech National Corpus

CxG Construction Grammar

DCPI Dictionary of Czech Phraseology and Idioms

ECDF empirical cumulative distribution function

nPVI normalized pairwise variability index

PLC Prague Linguistic Circle

VoIP Voice over Internet Protocol