Pronunciation of casual spoken Czech:

A quantitative survey

David Lukeš

Zuzana Komrsková

Marie Kopřivová

Petra Poukarová

Slavicorp 2018 | September 25th, 2018

CZECH NATIONAL CORPUS

Introduction

Overview

Point of departure

We have a new corpus of *casual spoken Czech* with *manual phonetic transcription* → **ORTOFON v1**

Goals

1. Here's some cool stuff you can do with it.
2. And by you, I mean *you*! 👉 😱 👈

Resources

Google: *korpus ortofon*

This presentation: https://trnka.korpus.cz/~lukes/slides/slavicorp2018/ortofon

Our wiki: https://wiki.korpus.cz/doku.php/cnk:ortofon

KonText query interface: https://kontext.korpus.cz/first_form?corpname=ortofon_v1

All the data via LINDAT:

- ELAN transcripts + audio: http://hdl.handle.net/11234/1-2579
- vertical: http://hdl.handle.net/11234/1-2580

Corpora of spoken Czech at the CNC I

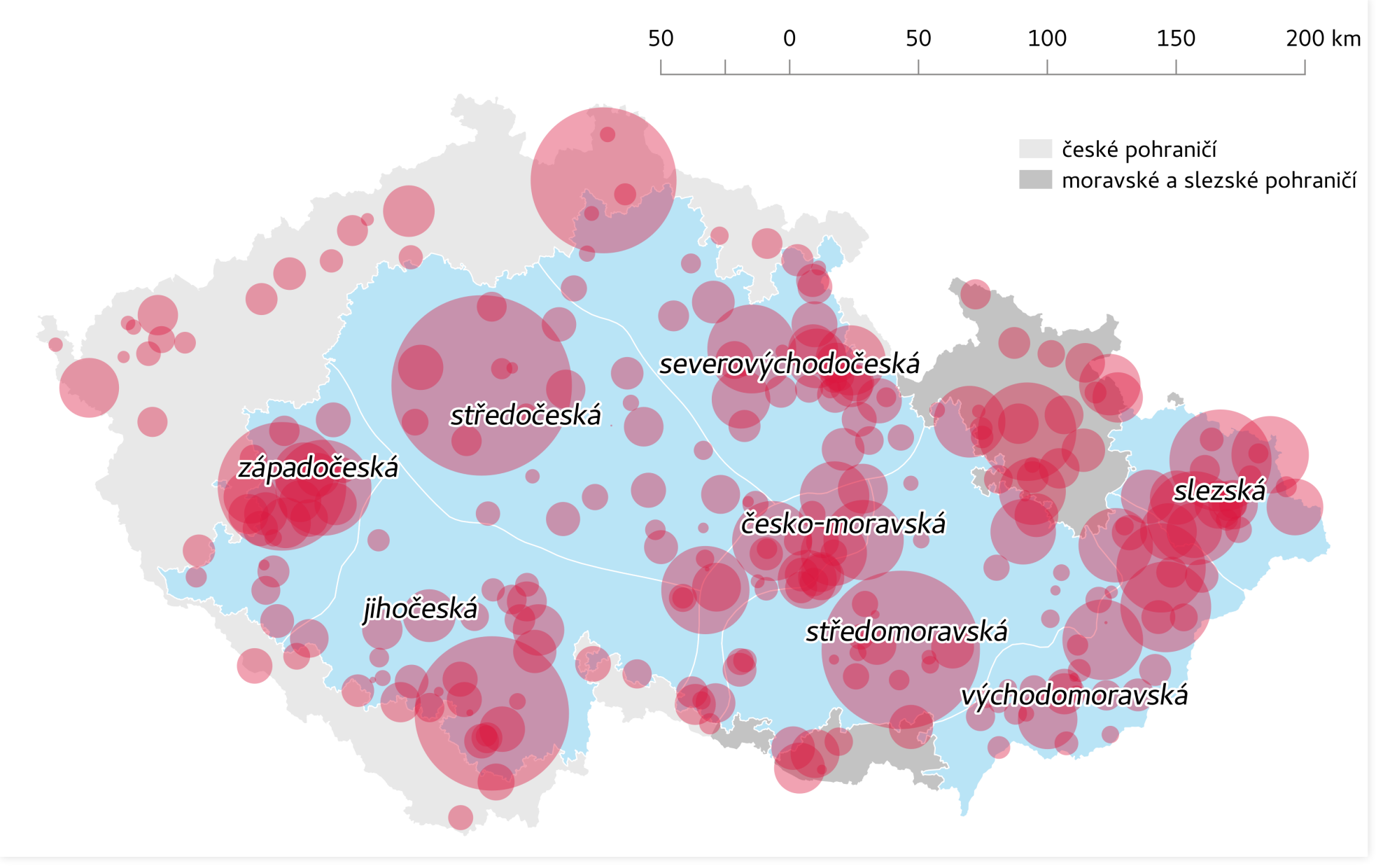
corpus	size	tagging	time span
ORTOFON	1M	✓	2012–2017
ORAL	5.4M	✓	2002–2011
↳ ORAL2013	2.8M	✗	2008–2011
↳ ORAL2008	1M	✗	2002–2007
↳ ORAL2006	1M	✗	2002–2006
BMK	490k	✗	1994–1999
PMK	675k	✗	1988–1996

Corpora of spoken Czech at the CNC II

corpus	size	tagging	time span
DIALEKT	100k	✓	1957–2015
LINDSEI_CZ	120k	✗	2012–2015
SCHOLA2010	790k	✗	2005–2008

About ORTOFON v1

Geographically



Some figures

# of ...	
... tokens	1,236,508
... tokens without puctuation, hesitations and interjections	1,014,786
... different word forms	65,294
... conversations recorded	332
... unique speakers	624
→ length of recordings [hh:mm:ss.ms]	102:41:14.247

Per-document metadata

- 12 pre-defined situation types
- year, month and location of recording
- relationship between speakers in recording
- gender mix of speakers in recording
- generation mix of speakers in recording
- … and more.

Per-speaker metadata: balancing

On the basis of the following metadata:

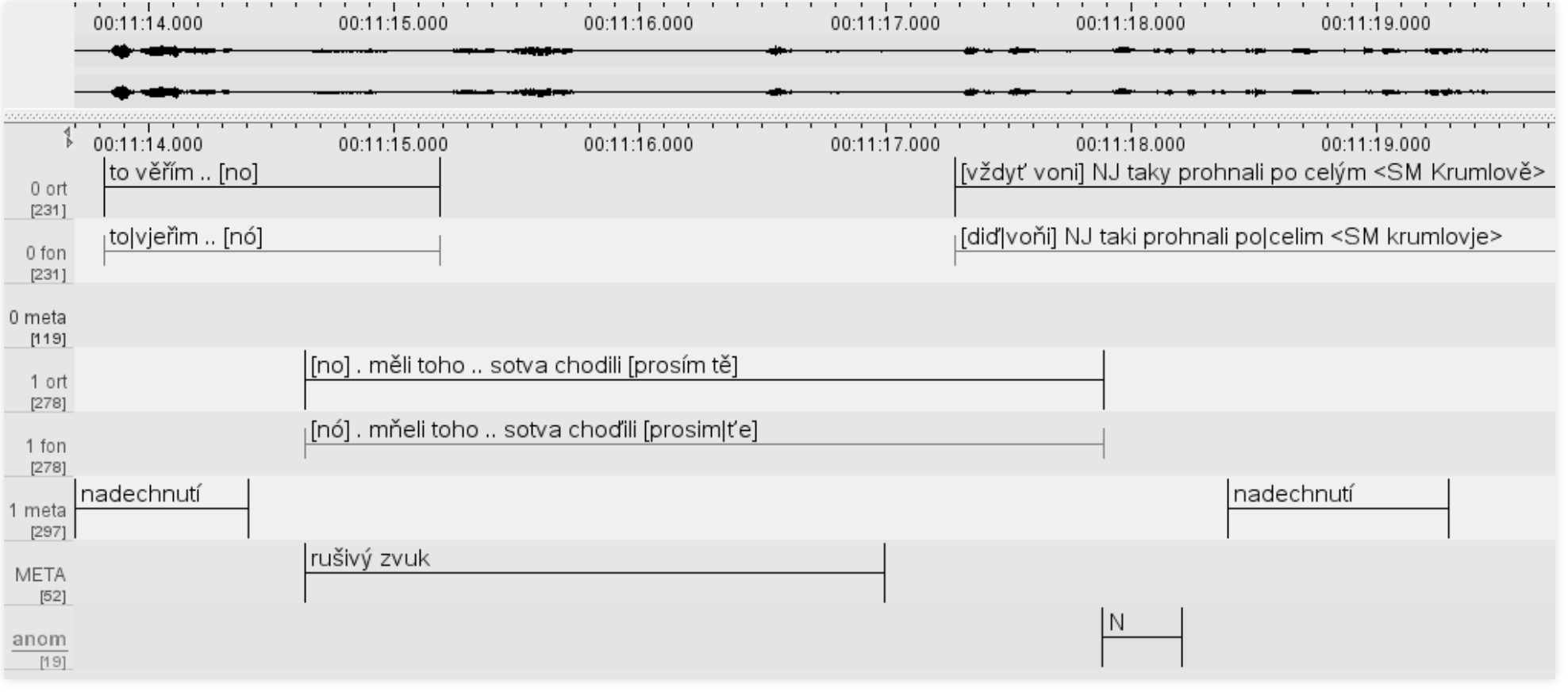
- *gender*: male × female
- *age*: under 35 × over 35
- *education*: tertiary × other
- *childhood dialect region of residence*: 10 regions

Resulting number of categories: $2 \times 2 \times 2 \times 10 = 80$

Ideally: **equal representation** of these 80 categories, at least 5 speakers per category.

→ Target number of words per category: $\frac{1\,000\,000}{80} = 12\,500$

Transcribed using ELAN



Accessible via KonText (but also LINDAT)

...	na ten převoz tý kočky ..
Kateřina_27	• a prostě fakt jako v září . se domluvím s tím doktorem v těch Bobnicích .
Kateřina_27	• s tím veterinářem u kterýho to stojí vo pětistovku míň než v Nymburce .. v Nymburce stojí kastrace kočky tisícovku
Josef_28	• (smrkání)
Josef_28	• no . <overlap> ve Veselí </overlap> taky asi pětistovku .. jsme o tom s babičkou mluvili

Phonetic transcription

- simplified phonetic transcription
  - use regular *Czech alphabet graphemes* as much as possible
  - no phonetic *diacritics*
- *stress group* boundaries
- *aligned* with basic transcript: tokens match $1:1$

Case studies

Assimilation of voicing in Czech

- final devoicing:

  <hra*d* (nom.), hra**d**u (gen.)> → [hra*t*, hra**d**u]

- regressive / anticipatory assimilation of voicing, even across word boundaries:

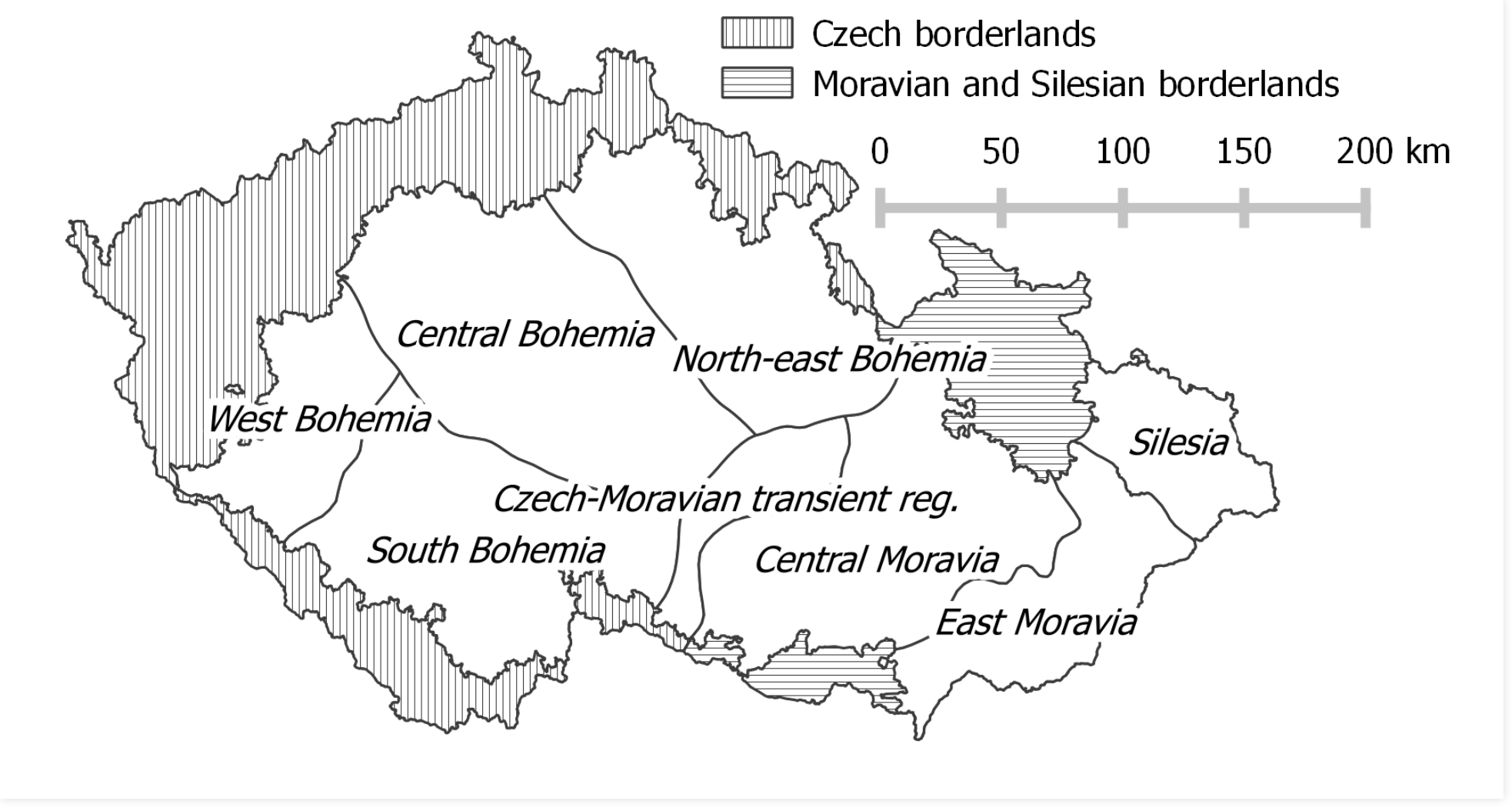
  <hra*d*, hra**d b**yl> → [hra*t*, hra**d b**il]

- triggered by voiced obstruents

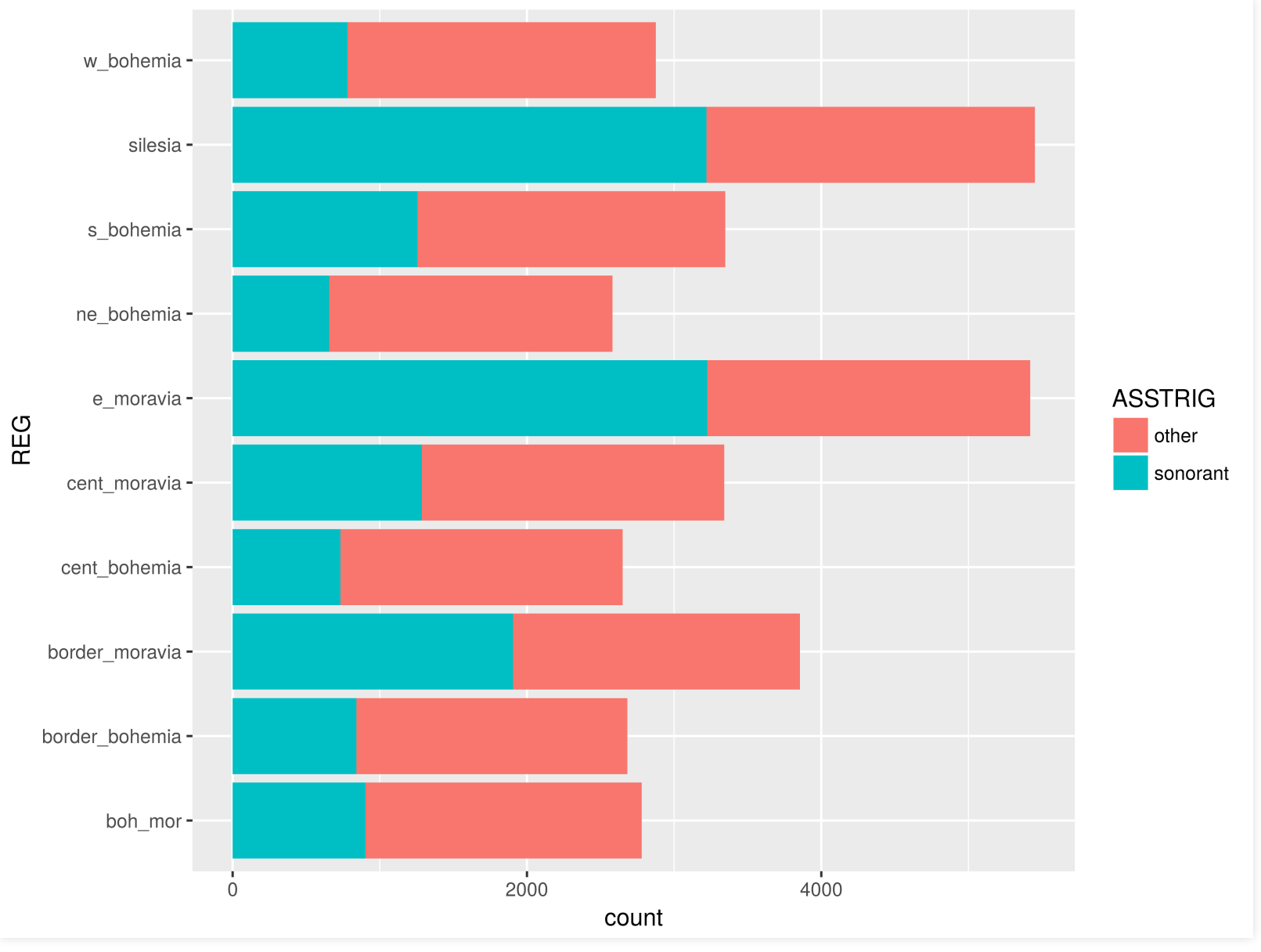
- in Moravia/Silesia (Eastern part of the country), also triggered by **sonorants** [r, l, m, n, j…]:

  <ta**k j**ako> → [ta**g j**ako]

Traditional dialect regions



Assimilation of voicing ~ Region of childhood residence



Which forms assimilate to sonorants?

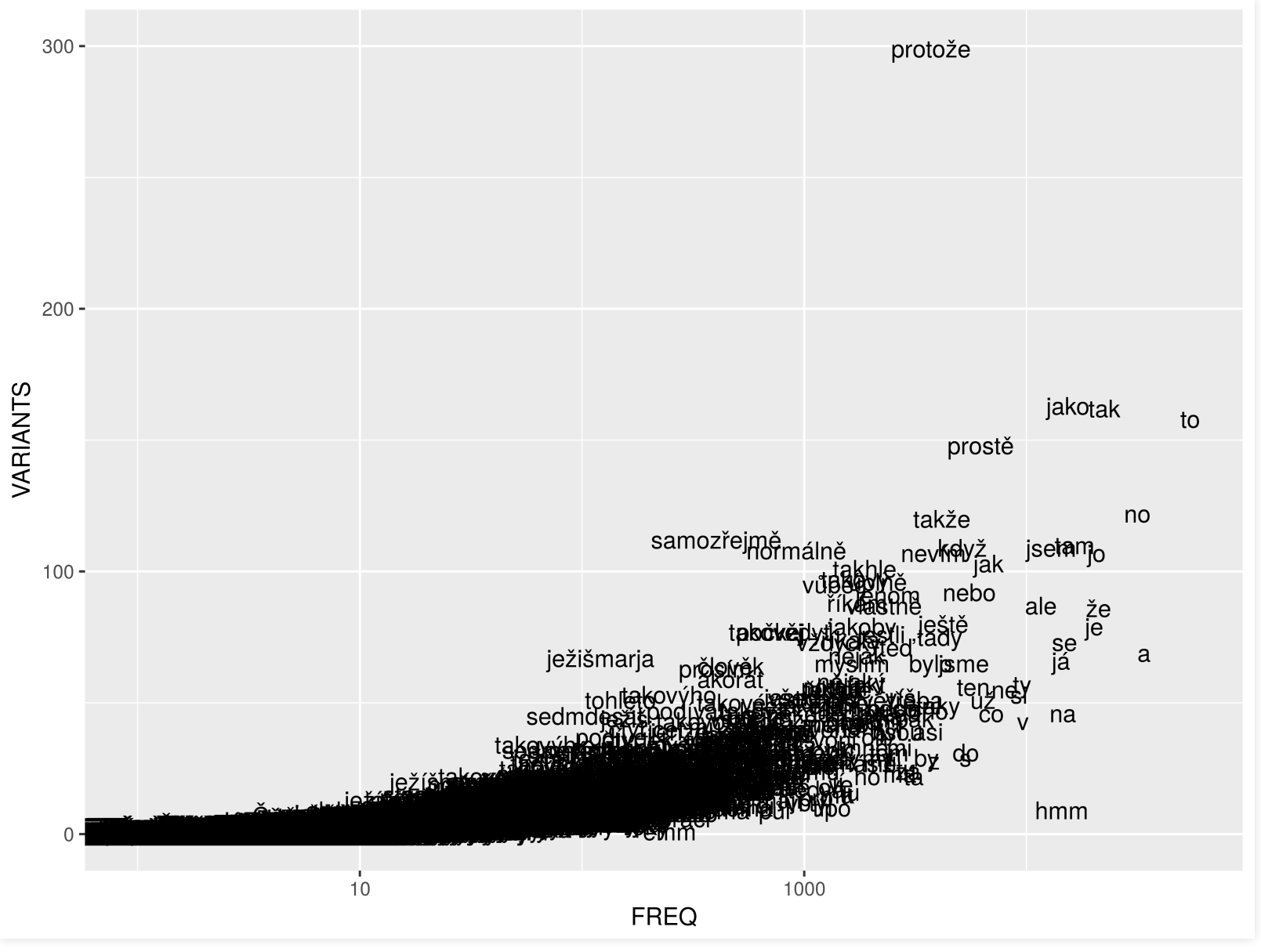
“Bohemia”

WORD	FREQ
tak	651
bych	348
už	239
těch	224
když	193
teď	186
vod	184
jak	169
vůbec	127
pak	122

“Moravia”

WORD	FREQ
tak	1288
už	483
jak	345
když	267
teď	190
fakt	181
vůbec	162
víš	155
bych	153
pak	145

of variants ~ Frequency ($\rho = 0.76$)



In KonText: protože

	Filter	fon	Freq
1	p / n	protože	660
2	p / n	prətəže	589
3	p / n	protəže	243
4	p / n	pətəže	235
5	p / n	prətože	188
6	p / n	pře	137
7	p / n	prtəže	104
8	p / n	bře	89

9	p / n	prze	79	
10	p / n	prəže	77	

In KonText: samozřejmě

	Filter	fon	Freq
1	p / n	samozřejmňe	57
2	p / n	samodřejmňe	50
3	p / n	samořejmňe	30
4	p / n	samozřejňe	21
5	p / n	samořeňe	15
6	p / n	samořejňe	15
7	p / n	samořemňe	14
8	p / n	samozřemňe	14
9	p / n	samodřemňe	11
10	p / n	samodřejňe	11

In KonText: normálně

	Filter	fon	Freq
1	p / n	normálňe	369
2	p / n	normáňe	87
3	p / n	normáe	64
4	p / n	normalňe	32
5	p / n	nomae	29
6	p / n	normae	28
7	p / n	nomáe	19
8	p / n	nərmálňe	18
9	p / n	nomáňe	17
10	p / n	nomaňe	17

Identify competition with entropy

	WORD	ENTROPY
1	ježišmarja	3.803729
2	samozřejmě	3.717063
3	protože	3.603883
4	sedmdesát	3.127680
5	takovýhle	3.110014
6	sedmnáct	3.096503
7	člověk	3.037660
8	ježíšmarjá	2.947005
9	šestnáct	2.927707
10	ježíš	2.883297
11	tohleto	2.880382
12	normálně	2.843373
13	povídám	2.782390

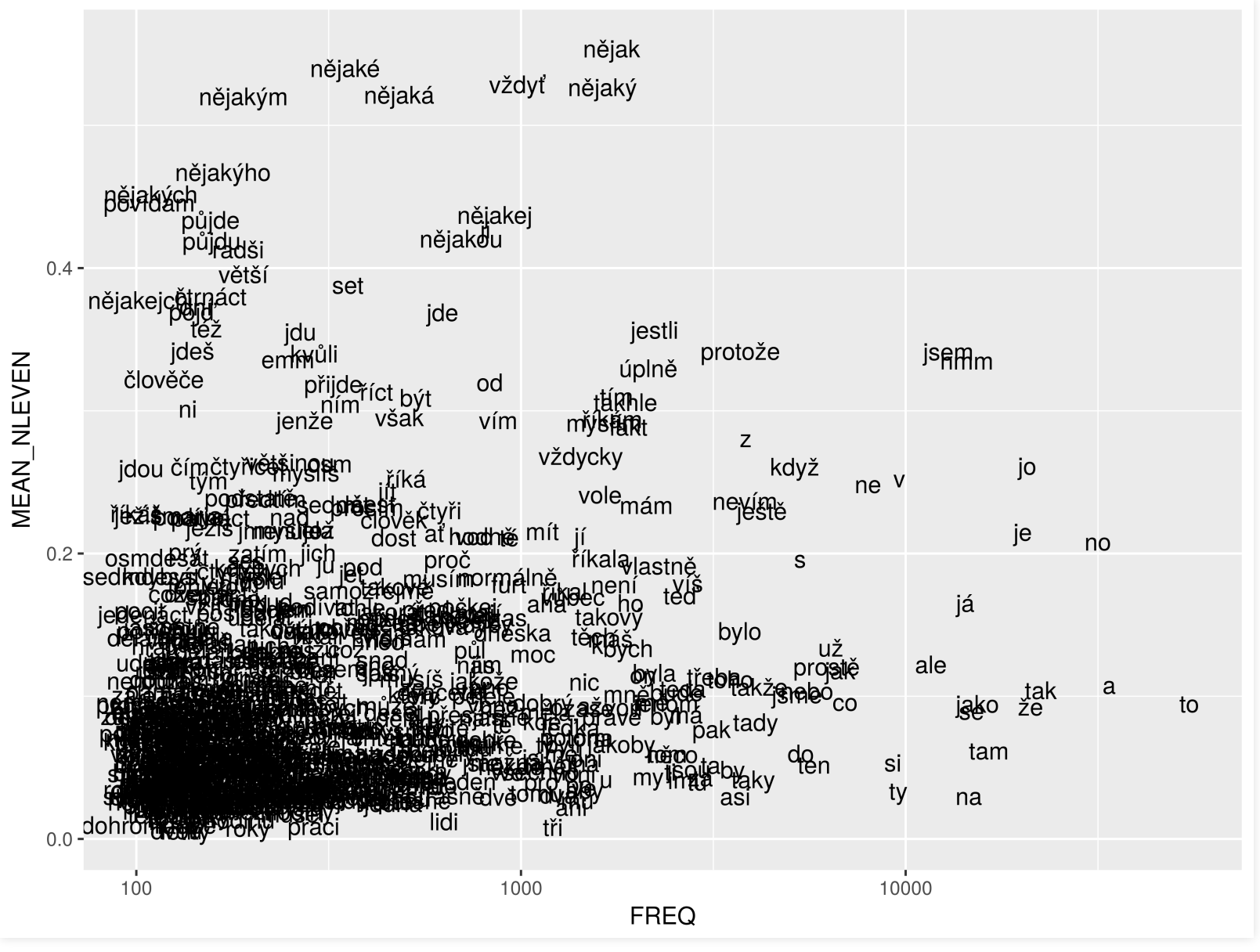
	WORD	ENTROPY
14	nějakého	2.752697
15	takového	2.682409
16	ježiš	2.680650
17	podívat	2.678791
18	tadyhle	2.676441
19	vůbec	2.671444
20	potřebovat	2.637769
21	čtyřicet	2.619200
22	myslíš	2.586492
23	přijít	2.574731
24	takovýho	2.573642
25	osmnáct	2.565948
26	ježíšmarja	2.523211

Formally reduced pronunciations

- motivated by:
  - frequency, length (across languages)
  - lexical effects
- edit distance between `abc` and `zbc`:
  - Levenshtein: 1
  - **normalized Levenshtein**: 0.33
- in practice, *deletion* and *substitution* (~ formal simplification) much more common than addition (**epenthesis**)

  → high normalized Levenshtein distance ~ high amount of simplification

Normalized Lev. dist. ~ Frequency



Instead of a conclusion…

Google: *korpus ortofon*

This presentation: https://trnka.korpus.cz/~lukes/slides/slavicorp2018/ortofon

Our wiki: https://wiki.korpus.cz/doku.php/cnk:ortofon

KonText query interface: https://kontext.korpus.cz/first_form?corpname=ortofon_v1

All the data via LINDAT:

- ELAN transcripts + audio: http://hdl.handle.net/11234/1-2579
- vertical: http://hdl.handle.net/11234/1-2580

Thank you for your attention!

Acknowledgments

This research was supported by the **Czech National Corpus** project (LM2015044) funded by the *Ministry of Education, Youth and Sports* of the Czech Republic within the framework of *Large Research, Development and Innovation Infrastructures*.

Slides:
https://trnka.korpus.cz/~lukes/slides/slavicorp2018/ortofon