Validita korpusu ORAL2013

Mgr. Jan Chromý, Ph.D.

Obsah

- korpus ORAL2013
- validita

ORAL2013

- cíl
  - „represent spontanneous spoken Czech in a sociolinguistically balanced way“ (Válková et al. 2012)

- charakteristiky
  - 835 nahrávek (17 471 minut; průměr = 21 minut) z let 2008–2011
  - 2 785 189 slovních tvarů
  - celkově 1297 zaznamenaných mluvčích

ORAL2013

- vyvážený
  - pohlaví
  - věk
  - vzdělání
  - region původu

- povaha nahrávek
  - neformální rozhovory
  - většinou skryté nahrávání

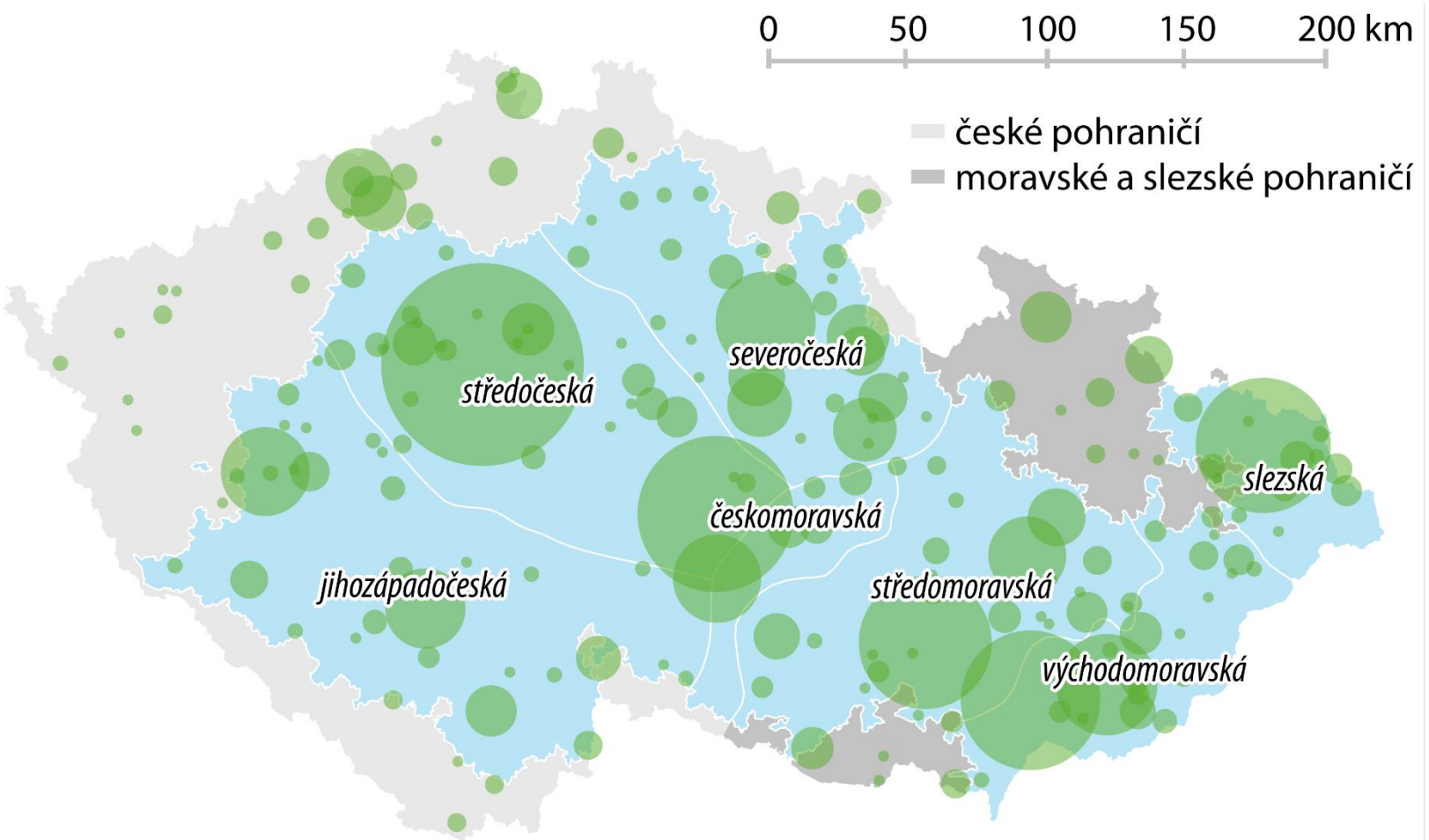
ORAL2013: problémy

- stavba vzorku a reprezentativnost
- sociolingvistické údaje

Problémy stavby vzorku

- pokrytí území
- „vyváženost“

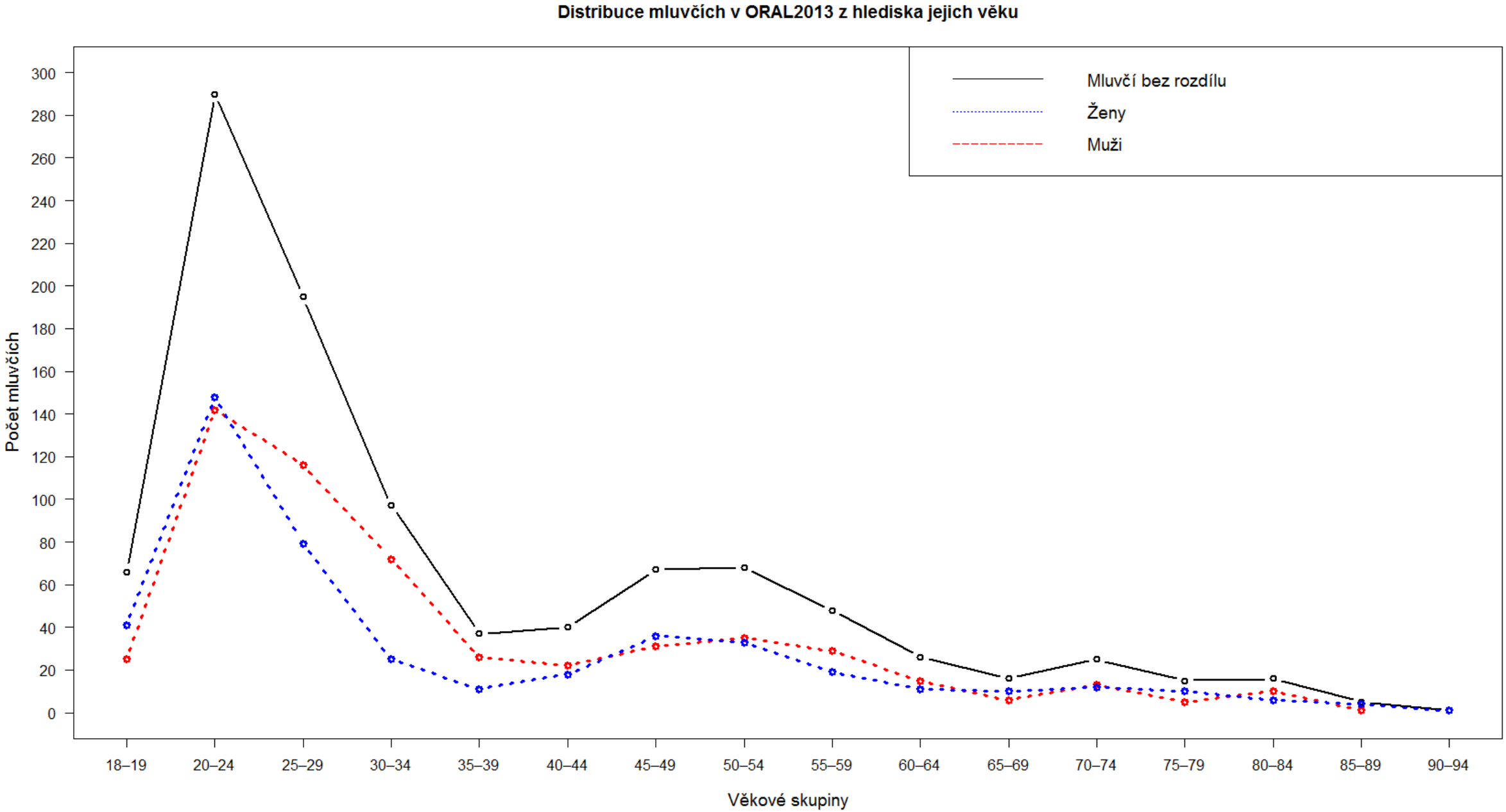
ORAL2013: pokrytí



převzato z http://wiki.korpus.cz/lib/exe/detail.php/cnk:o13.png?id=cnk%3Aoral2013

ORAL2013: vyváženost

- věk
  - < 35 let vs. ≥ 35 let
  - kontinuum 18–93 let rozděleno nerovnoměrně
- vhodnější by byla životní stadia
  - adolescence
  - raná dospělost
  - rodičovství
  - důchod


Distribuce mluvčích v ORAL2013 z hlediska jejich věku
Mluvčí bez rozdílu
Ženy
Muži
Počet mluvčích
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
18–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
75–79
80–84
85–89
90–94
Věkové skupiny

Problémy s údaji

- málo demografických informací
  - věk
  - vzdělání
  - místo původu
  - vztah k nahrávajícímu

- problém místa původu
  - pozdější mobilita?

Místo původu

- devět oblastí, žádná bližší specifikace
  - jihozápadočeská
  - pohraničí moravské
  - pohraničí české
  - severovýchodočeská
  - slezská
  - středomoravská
  - středočeská
  - východomoravská
  - česko-moravská

Validita

- Nakolik můžeme prostřednictvím korpusu ORAL2013 zkoumat současnou spontánní mluvenou češtinu?

- souběžná validita
  - dva způsoby zkoumání téhož by měly vykazovat stejné výsledky

Souběžná validita

- srovnání dvou zdrojů dat
  - data z grantu GAČR
  - analogická data z ORAL2013

- vzorek
  - mluvčí z Prahy 20–30 let
  - mluvčí z Českých Budějovic 19–30 let
- srovnávané jevy
  - protetické /v/
  - kondicionálový tvar 1. os. pomocného slovesa být

Grant GAČR 13-12973P (2013–2015)

- Sociolingvistická analýza užívání protetického /v/ v Čechách (SAUP)
- pět zkoumaných měst
  - 10 žen a 10 mužů ve věku cca 20–30 let
  - 10 žen a 10 mužů ve věku cca 60–70 let
- respondenti
  - narodil/a se a žije v daném místě
  - alespoň jeden z rodičů z daného místa, druhý z Čech v užším slova smyslu
- hodinové rozhovory
  - neformální

Použité datové soubory

- částečně rozdílná povaha dat

	Mluvčí kondicionál	Výskyty kondicionál	Mluvčí (v)	Výskyty (v)
SAUP: Praha	18	697	18	3359
ORAL: STČ	114	868	127	5495
SAUP: ČB	20	732	20	4780
ORAL: JZČ	57	534	67	3550

Kondicionál

- výchozí předpoklad
  - v 1sg preference bych, v 1pl preference bysme

- kódování
  - varianty bych, kdybych, abych nejsou rozlišovány
  - dvě hodnoty: spisovný vs. nespisovný tvar

Kondicionál: výsledky

	bych	*bysem*	*bychom*	*bysme*
SAUP: Praha	631 (98,9 %)	7 (1,1 %)	1 (1,69 %)	58 (98,31 %)
ORAL: STČ	674 (98,25 %)	12 (1,75 %)	13 (7,14 %)	169 (92,86 %)
SAUP: ČB	637 (98,61 %)	9 (1,39 %)	8 (9,3 %)	78 (90,7 %)
ORAL: JZČ	406 (97,83 %)	9 (2,17 %)	0 (0 %)	119 (100 %)

- rozdíl
  - pouze 1pl SAUP: ČB vs. ORAL: JZčeská

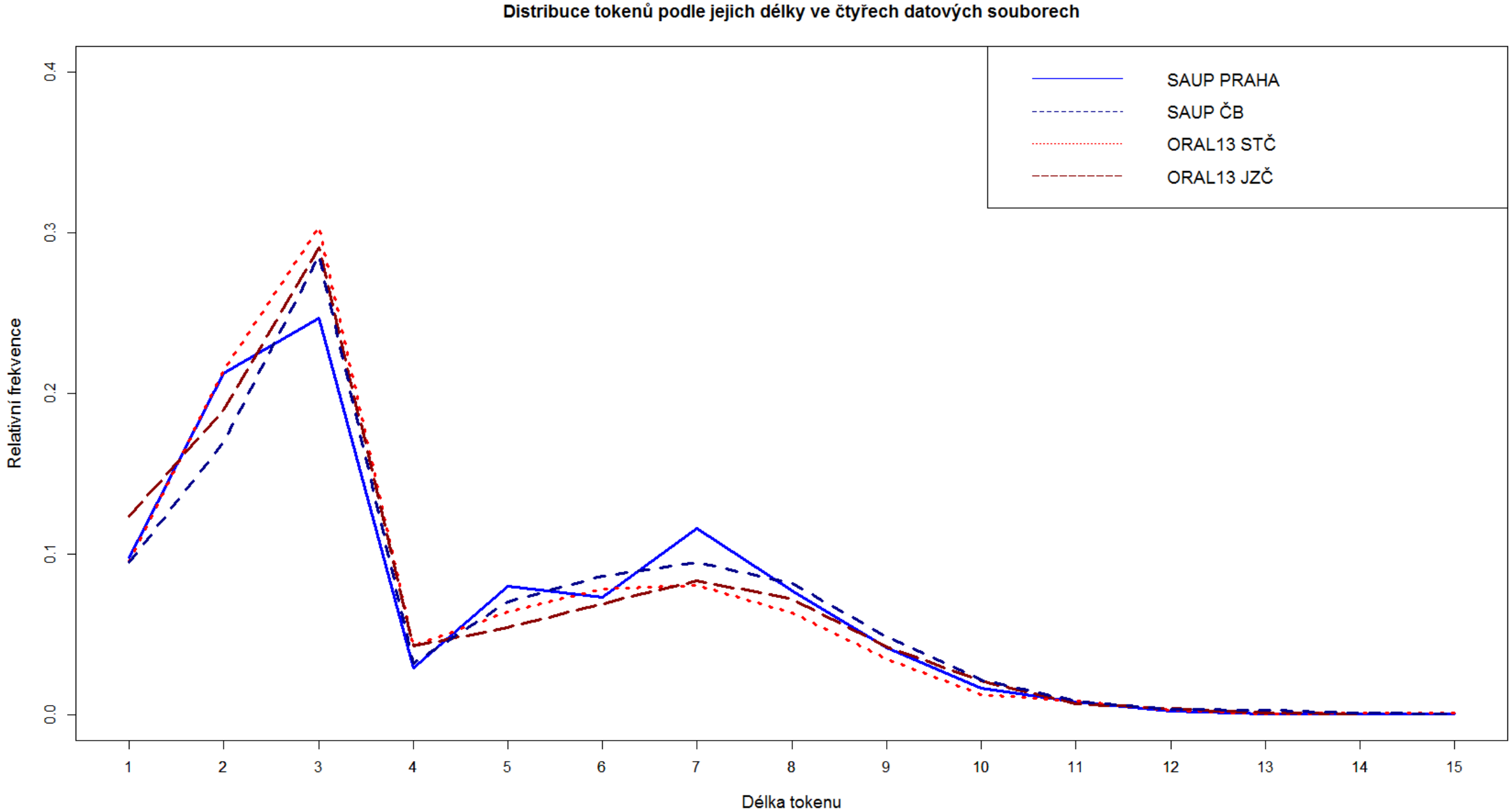
Průběžné závěry: Omezení dat

- SAUP
  - všech 8 výskytů *bychom* je u jednoho mluvčího
  - problém dopadu jednotlivých mluvčích u malého vzorku

- ORAL2013
  - krátké rozhovory
  - 1pl alespoň jeden výskyt pouze u 56 mluvčích (ze 114 v STČ) a u 34 mluvčích (z 57 v JZČ)
  - 31 výskytů 1pl (cca 25 %) u jednoho mluvčího v JZČ

Protetické v-

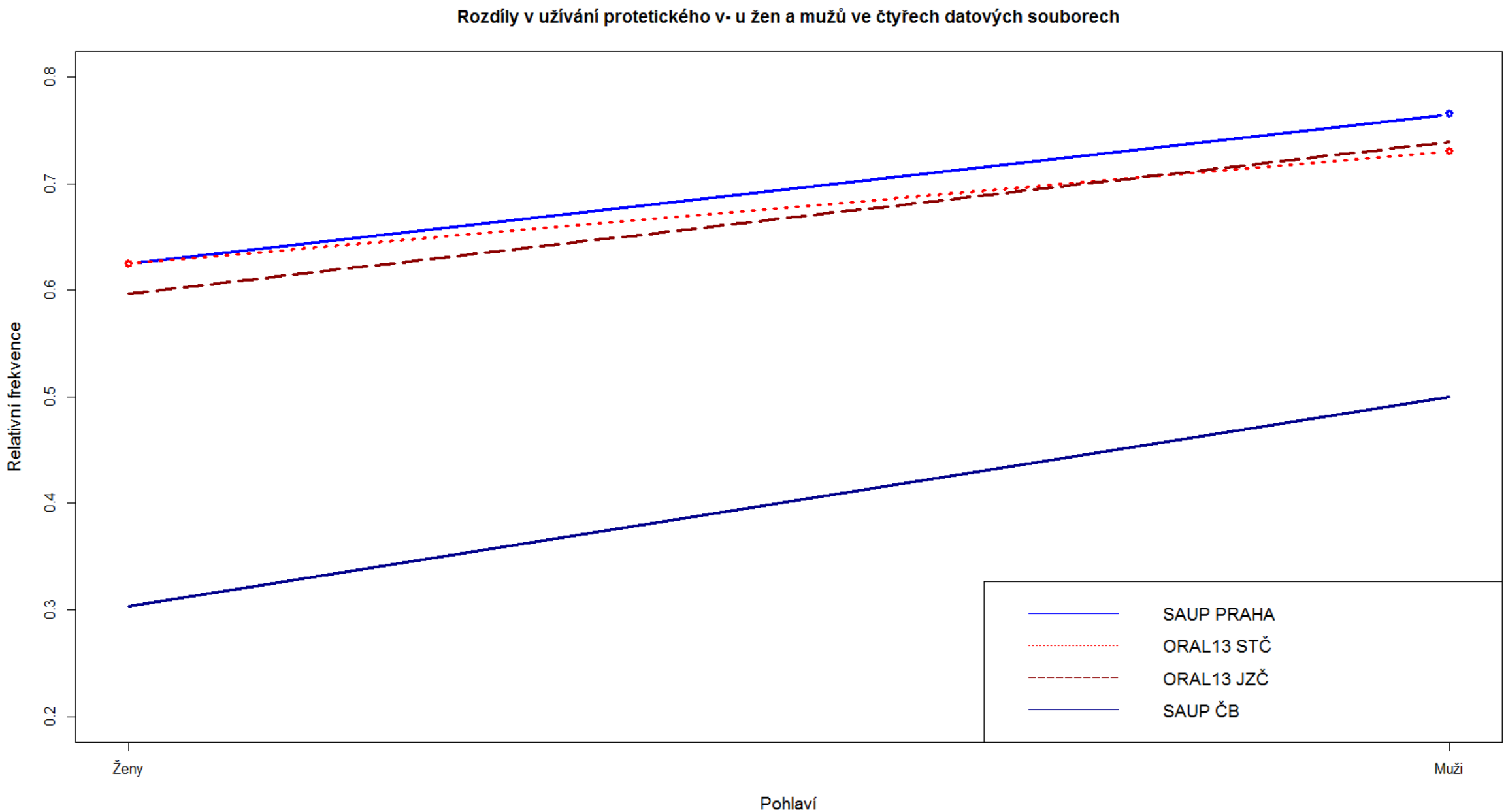
- Chromý (2015)
  - užívání je spoluurčováno řadou faktorů
- obtížnější analýza
  - jsou si datové soubory základně podobné?


Distribuce tokenů podle jejich délky ve čtyřech datových souborech
SAUP PRAHA
SAUP ČB
ORAL13 STČ
ORAL13 JZČ
Relativní frekvence
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Délka tokenu

Protetické v-: výsledky

	ženy		muži	
	/o/	*/vo/*	*/o/*	*/vo/*
SAUP: Praha	689 (37.49 %)	1149 (62.51 %)	341 (23.45 %)	1113 (76.55 %)
ORAL: STČ	1253 (37.49 %)	2089 (62.51 %)	580 (26.94 %)	1573 (73.06 %)
SAUP: ČB	1778 (69.64 %)	775 (30.36 %)	1114 (50.02 %)	1113 (49.98 %)
ORAL: JZČ	643 (40.31 %)	952 (59.69 %)	510 (26.09 %)	1445 (73.91 %)

- rozdíl
  - SAUP: ČB vs. ORAL: JZČ


Rozdíly v užívání protetického v- u žen a mužů ve čtyřech datových souborech
Relativní frekvence
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Ženy
Muži
Pohlaví
SAUP PRAHA
ORAL13 STČ
ORAL13 JZČ
SAUP ČB

Důvody rozdílu

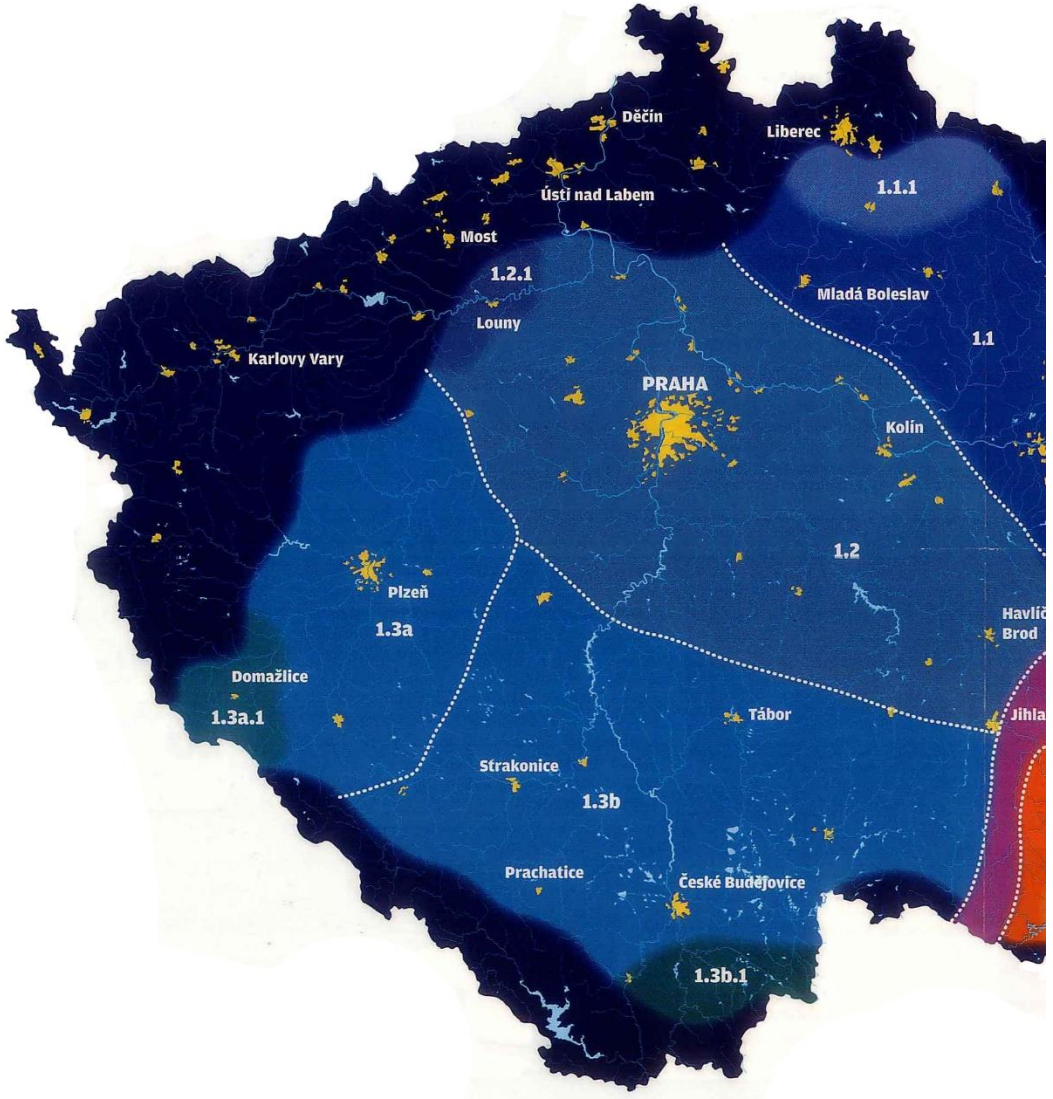
- vzorek SAUP
  - mluvčí, kteří se nevyjadřují spontánně
- vzorek ORAL2013: JZČ
  - silně heterogenní
- další faktory
  - ?

Nespontánnost v SAUP

- nepravděpodobné
  - viz výsledky kondicionálu
  - zcela běžně další obecně české varianty (ý > ej, é > ý, -ama apod.)

Vzorek ORAL: JZČ

- heterogennost
  - jihozápadočeská oblast
  - Plzeň vs. České Budějovice vs. …
- České Budějovice na hranici doudlebského výběžku
  - historicky bez protetického v-



převzato z http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-dialektologicke/publikacni-cinnost/obalky/mapa-nareci.jpg a upraveno

Závěr

- středočeská oblast
  - zdá se validní pro pražské mluvčí
  - žádné rozdíly v užívání kondicionálových tvarů a protetického v-
- jihozápadočeská oblast
  - příliš velká
  - různé jevy mohou být používány různě v různých místech

Závěr

- rozdíly mezi mluvčími
  - disproporce mezi objemem dat od různých mluvčích v ORAL2013
  - v případě malého vzorku velký problém

- potřeba dodatečných informací
  - mobilita mluvčích