Registrová variabilita českého internetu: možnosti a meze diskrétní kategorizace textů na základě výsledků multidimenzionální analýzy

Jan Henyš · jan.henys@ff.cuni.cz · Žďárek 2019 · listopad 2019

Multidimenzionální analýza registrové variability

- vytvoření korpusu
- vytvoření seznamu variujících jazykových rysů
- změření relativizovaných hodnot rysů pro texty v korpusu
- faktorová analýza
- změření faktorových skóre pro texty v korpusu
- interpretace dimenzí

Výsledné dimenze (Cvrček et al. 2016)

- 1. dynamický (+) × statický (-)
- 2. spontánní (+) × připravený (-)
- 3. vyšší (+) × nižší (-) stupeň koheze
- 4. polytematický (+) × monotematický (-)
- 5. vyšší (+) × nižší (-) míra adresnosti
- 6. obecný (+) × konkrétní
- 7. prospektivní (+) × retrospektivní (-)
- 8. postojový (+) × faktuální (-)

Cvrček et al. (2016) × Henyš (2019)

Korpus Koditex

↓

Seznam rysů

↓

Hodnoty rysů

↓

Faktorová analýza

↓

Naměření faktorových skóre

↓

Interpretace dimenzí → Naměření f. skóre pro Araneum → man. anotace → definice subregistrů

↓ ↓

Popis registrové variability češtiny Popis registrové variability internetových textů

Vzorek Websample

- Distribucí délek textů kopíruje Koditex
- Subkorpus o 1000 textech
- Zdrojem webový korpus řady Aranea

Dostupné informace

- **ID: identifikační řetězec**
- LEN: kategorie podle délky textu
- CRAWLDATE: datum stažení z internetu
- **URL: přesná adresa dokumentu**
  - např. http://www.cinske-zbozi.cz/cinska-cesta-zdravi/bylinky-lecive-byliny-bylinne-leky/351-bylinky-jako-lek-a-zdroj-vnitrni-energie

Dostupné informace

- AREF: adresa webu (doména III. řádu, subdoména)
  - např. cinske-zbozi.cz
- LANGDIFF: metrika pro zjištění jinojazyčnosti
- TLD: doména I. řádu
  - např .cz, .com

Přidané informace

- URLLIVE: on-line dostupnost textu
- SUBREG: subregistrová kategorizace
- NOTE: poznámky
- DEL: označení nevyhovujících textů
- GLS1…GLS8: faktorová skóre dimenzí variability

Postup práce

- Přístup k textu z korpusu Websample přes rozhraní KonText
- Kontrola přístupu přes prohlížeč
  - zaznamenání nedostupnosti při chybách 404, 403, 500 apod., blokování antivirem
  - kontrola shody textu v korpusu a prohlížeči
- Subregistrová klasifikace
  - četba textu
  - subregistrové markery
  - URL a mimotextové charakteristiky (v případě dostupnosti)

Subregistrová kategorizace

- Zdroj: Biber–Egbert: Register Variation on the Searchable Web: A Multi-Dimensional Analysis (2016)
- 8 obecných registrů a 27 subregistrů
- revize subregistrů ve dvou fázích
  - slučování nedostatečně distinktivních registrů, odstraňování nevyhovujících kategorií, úprava definic, přidání registrů
  - druhá revize po pilotáži (10 % anotovaného vzorku)

Subregistrová kategorizace

1. Poradenský text (ADV)
2. Popis s účelem prodeje (DES)
3. Popis předmětu (DET) + popis osoby
4. Diskusní fórum (DIS)
5. Nejčastěji kladené dotazy (FAQ)
6. Encyklopedické články (ENC) + vědecké články/texty

Subregistrová kategorizace

7. Umělecká próza (FIC) + písňové texty a poezie
8. Formální projev (FOR)
9. Historický článek (HIS)
10. Hobby a volný čas (HOB)
11. Návod typu „jak na to“, instruktáž (HOW)
12. Informační text (INF)

Subregistrová kategorizace

13. Interview (INT)
14. Právní text (LEG)
15. Zpravodajský text/blog (NEW)
16. Názorový blog (PEO)
17. Osobní blog (PER)
18. Fórum typu Q-A (QAF)

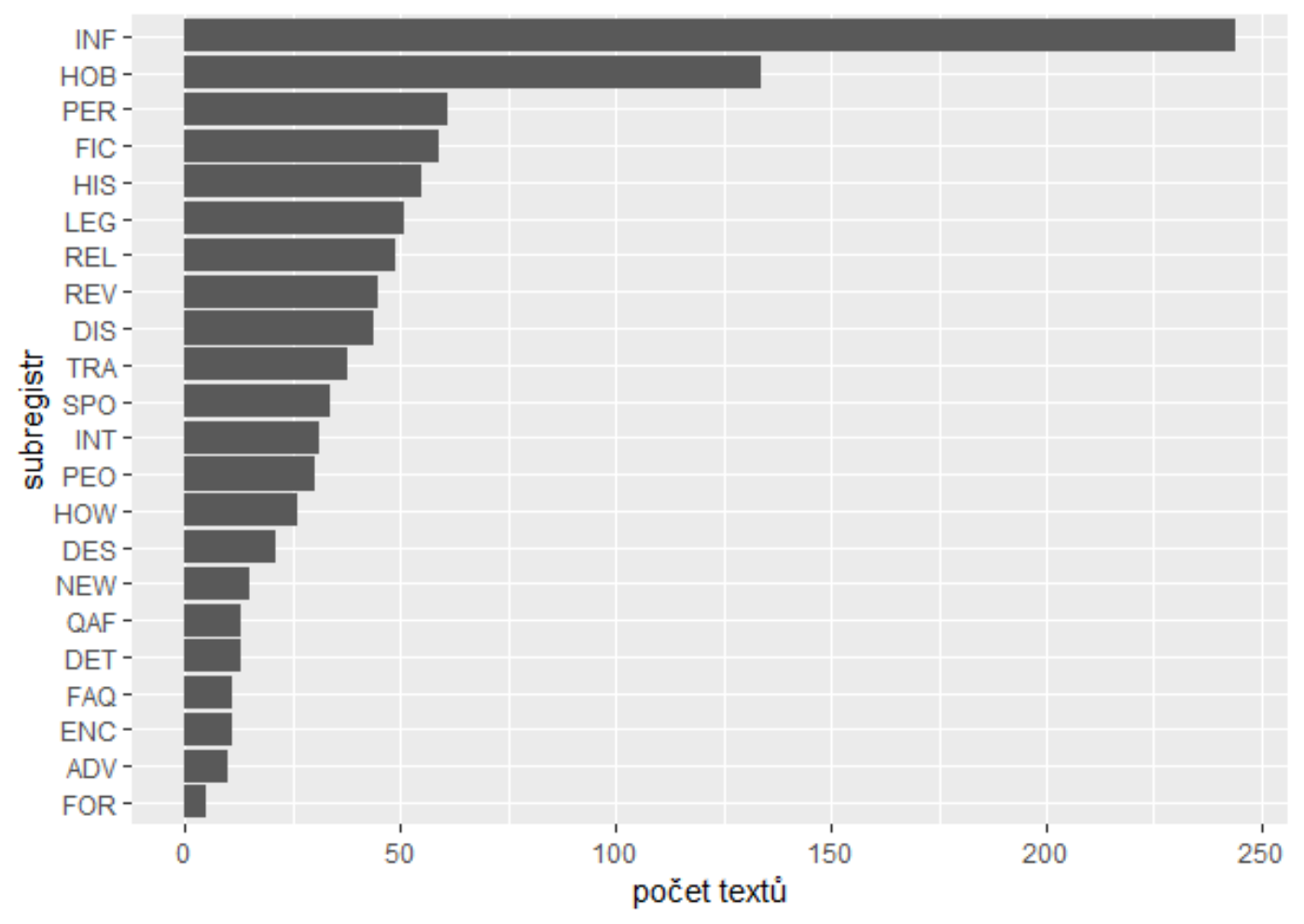
Subregistrová kategorizace

19. Náboženský blog/kázání (REL) + ezoterie a parapsychologie
20. Recenze (REV)
21. Sportovní reportáž/článek (SPO)
22. Cestovatelský blog (TRA)

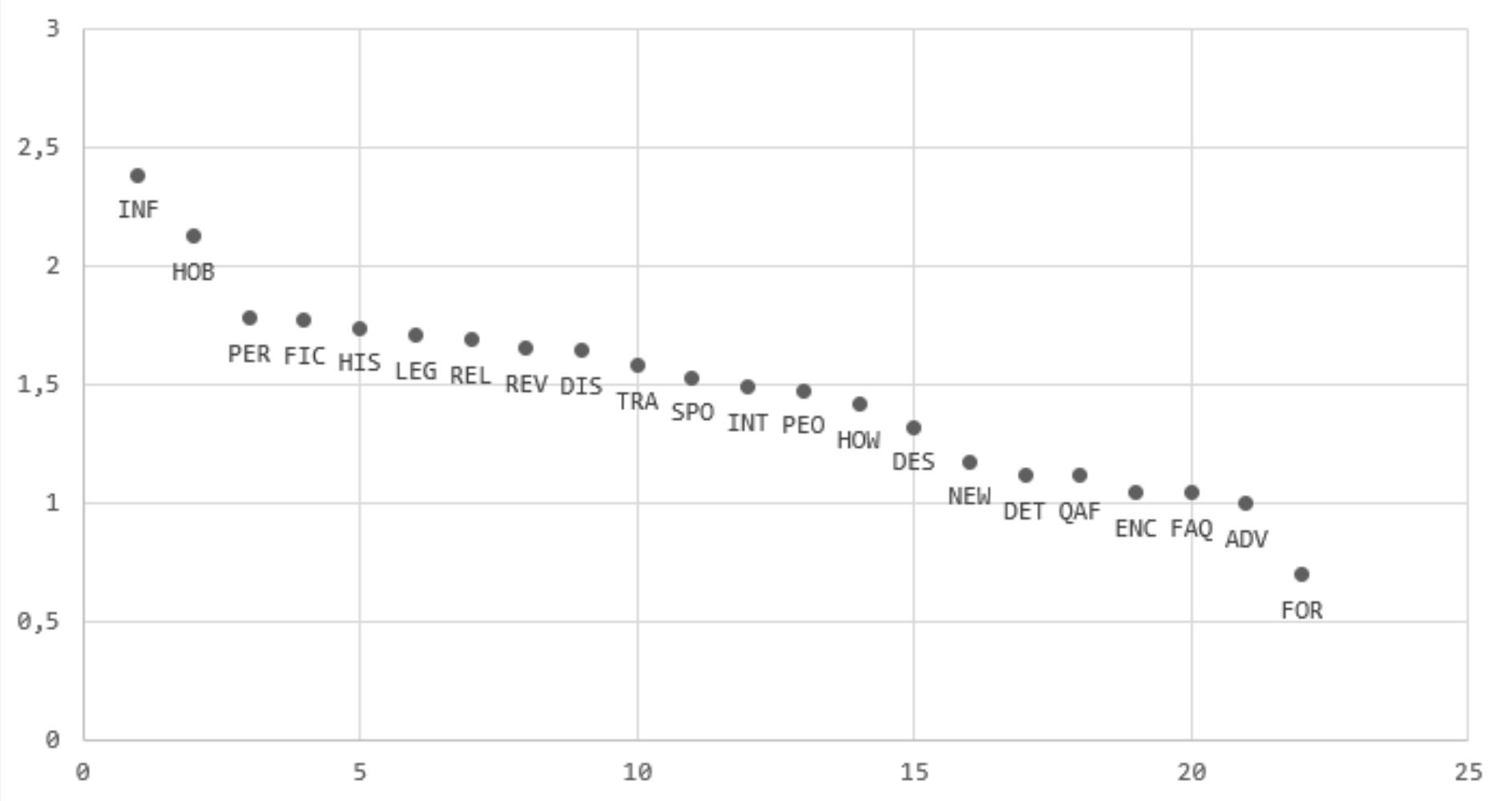
22 subregistrů

- Poradenský text
- Popis s účelem prodeje
- Popis předmětu
  - + popis osoby
- Diskusní fórum
- Encyklopedický článek
  - + vědecký článek
- Nejčastěji kladené dotazy (FAQs)
- Umělecká próza
  - + písně a poezie
- Formální projev
- Historický článek
- Hobby a volný čas
- Návod typu jak na to, instruktáž
- Informační text
- Interview
- Právní text
- Zpravodajský článek/blog
- Názorový blog
- Osobní blog
- Fórum typu question-answer
- Náboženský blog / kázání
  - + ezoterie a parapsychologie
- Recenze
- Sportovní reportáž
- Cestovatelský blog

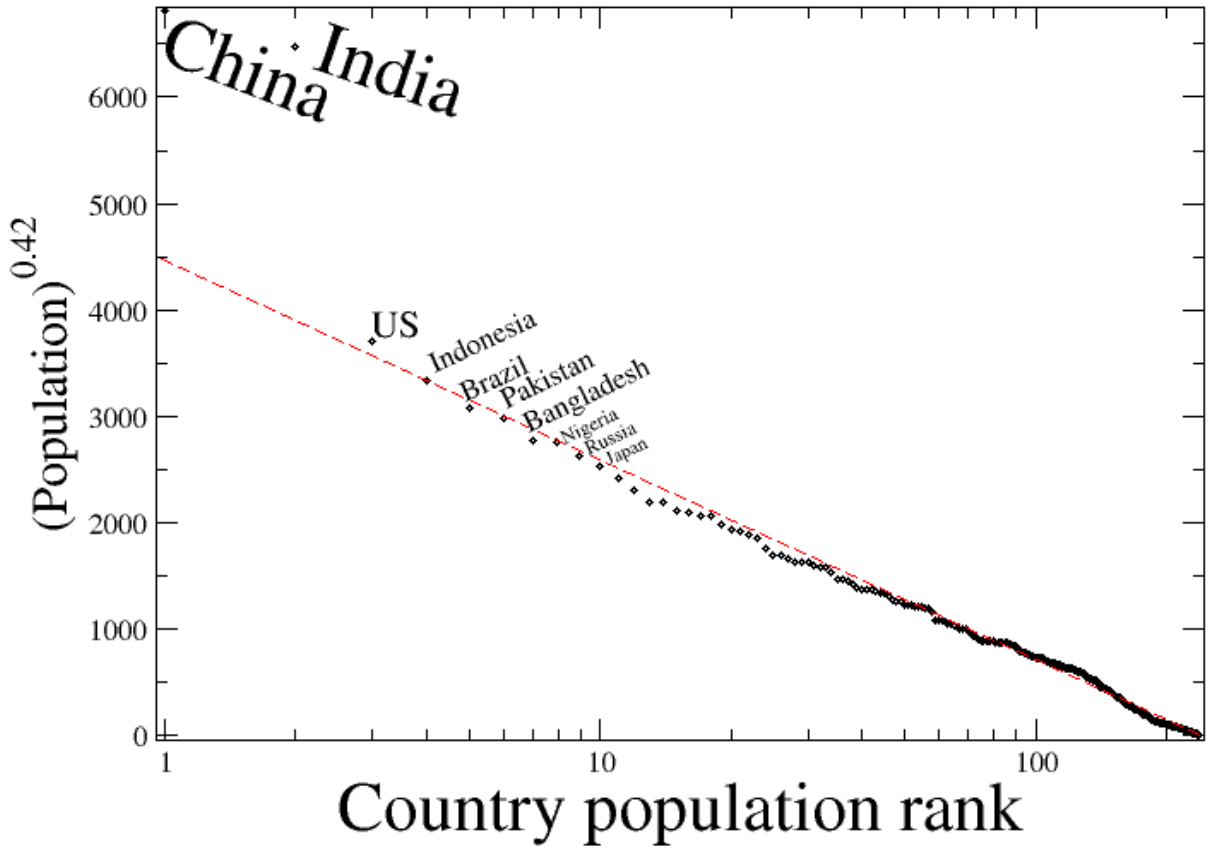
Distribuce textů v subregistrech



Distribuce textů v subregistrech



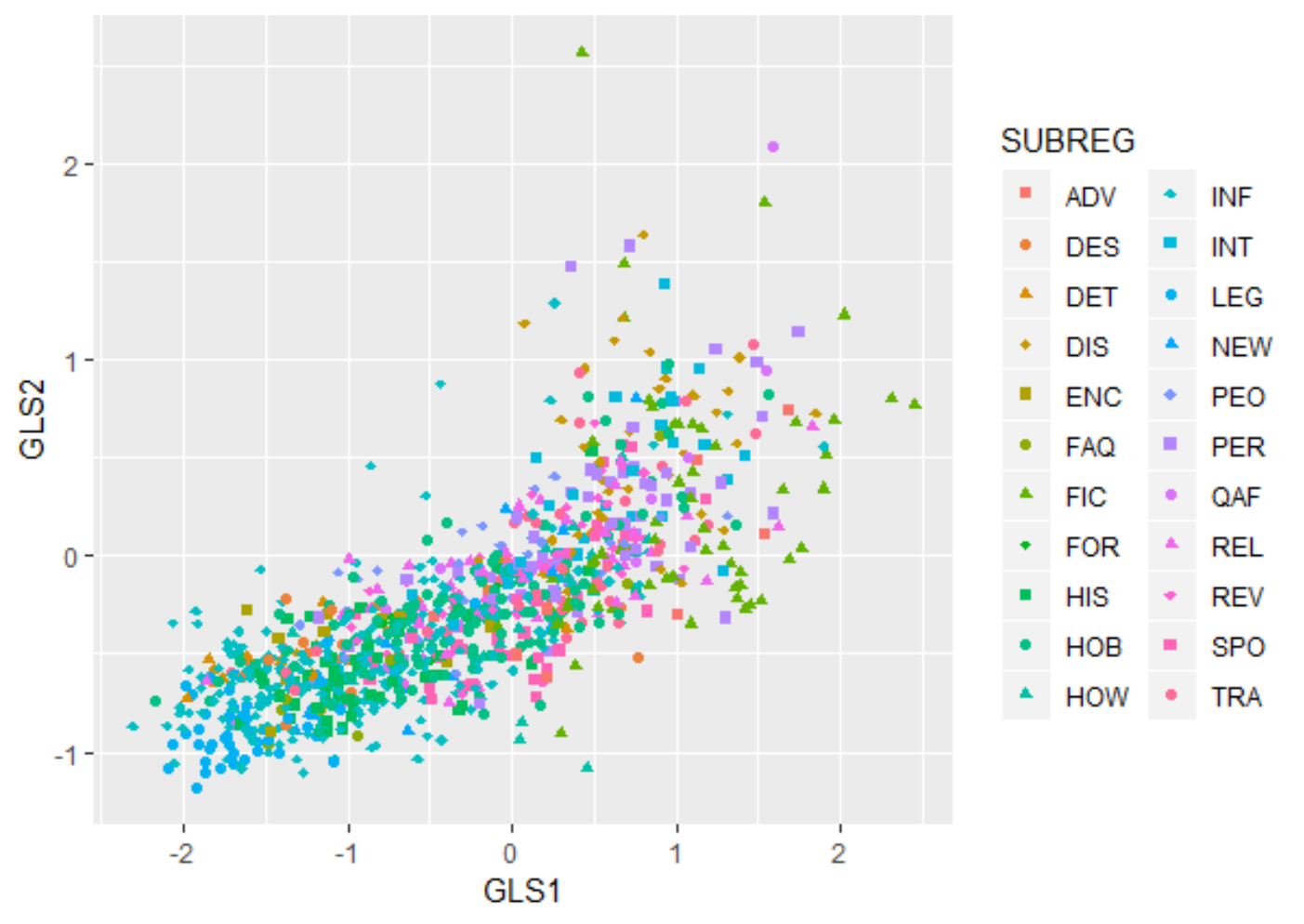
Distribuce populace ve státech - king effect



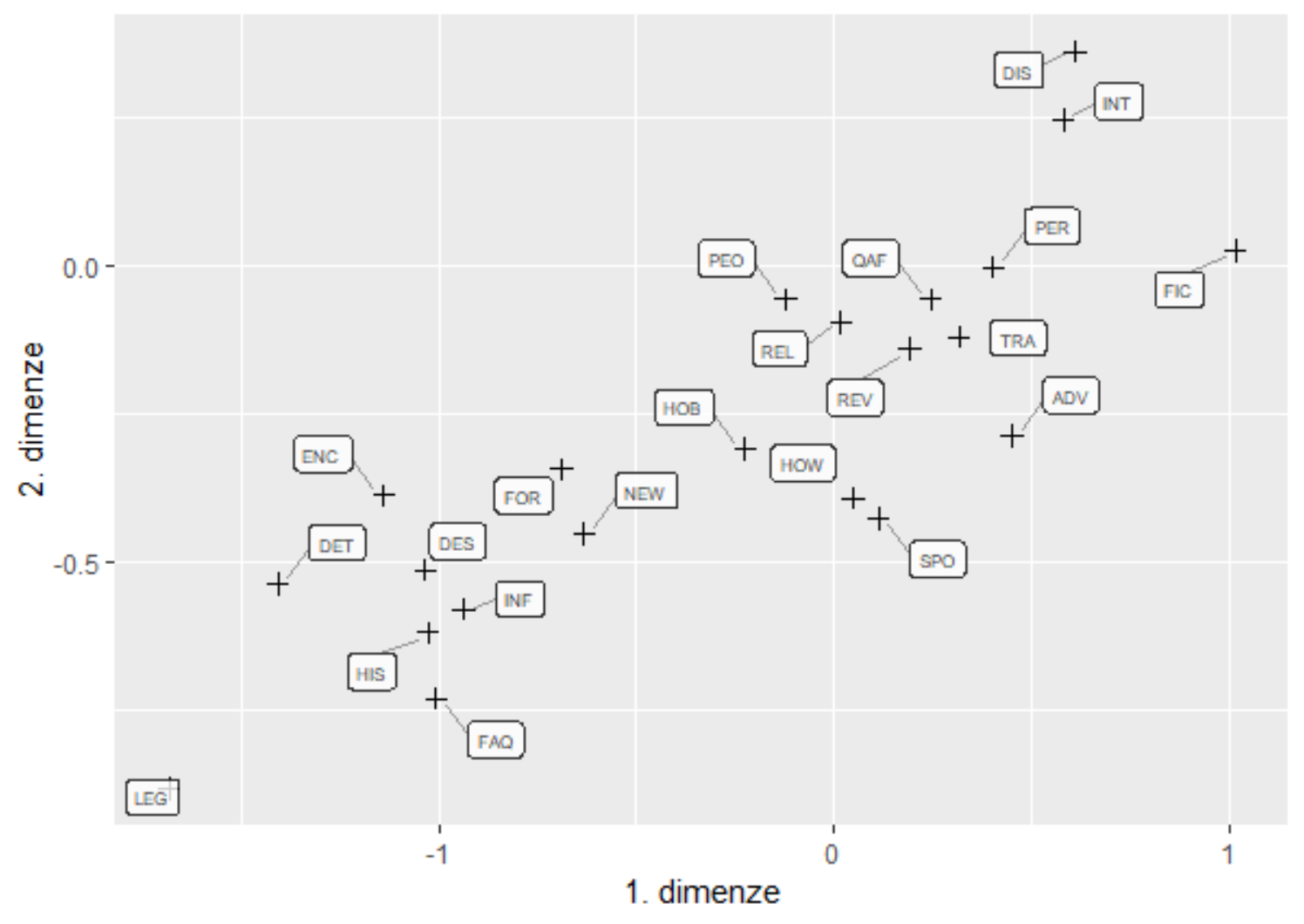
Výsledky MDA

- 22 diskrétních kategorií
- 8 dimenzí
- **40,5 %** textů dostupných on-line (červenec 2019)
- každý z textů je možné reprezentovat bodem v multidimenzionálním prostoru

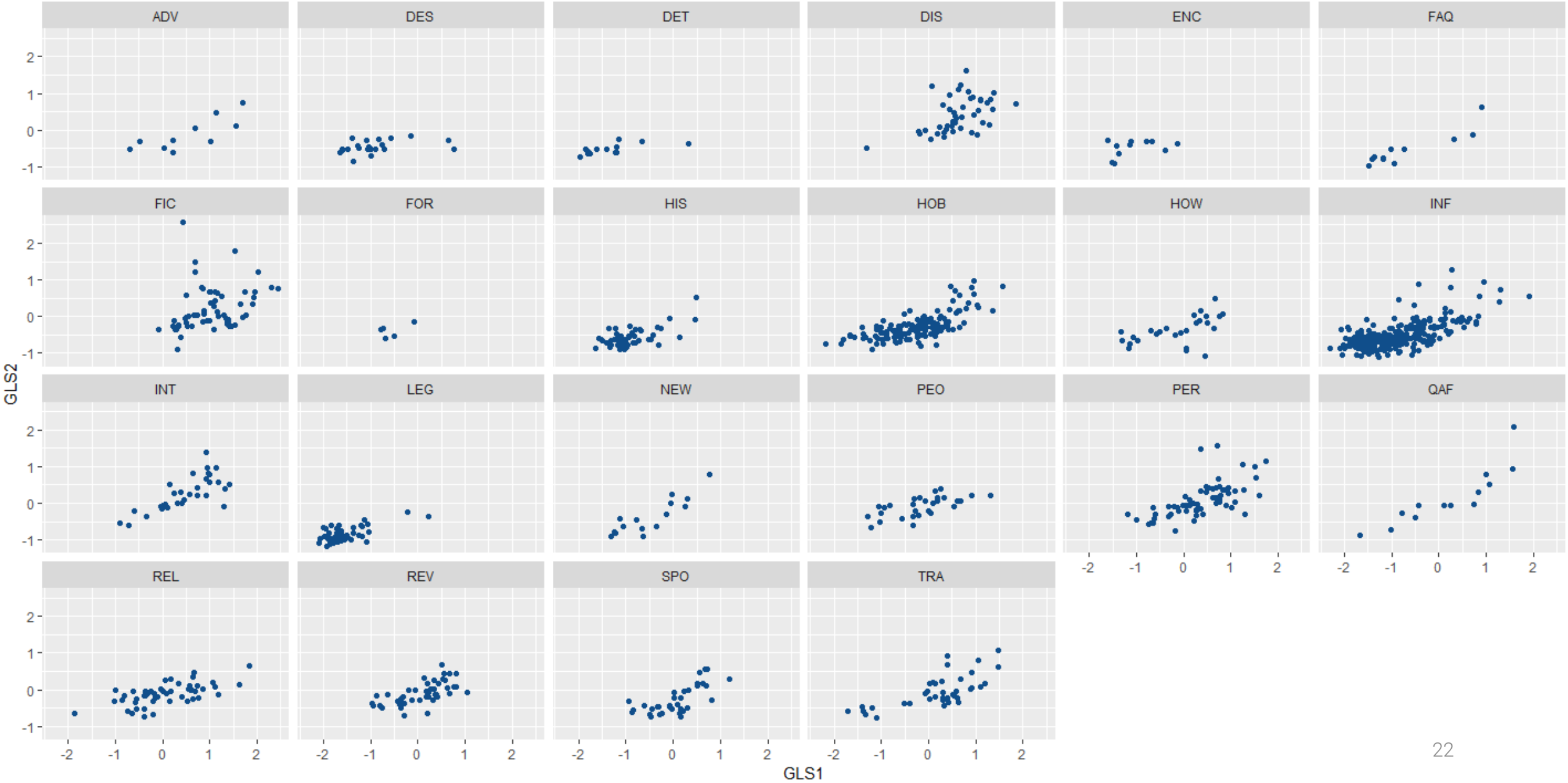
I (dyn+stat-) × II (spont+prep-)



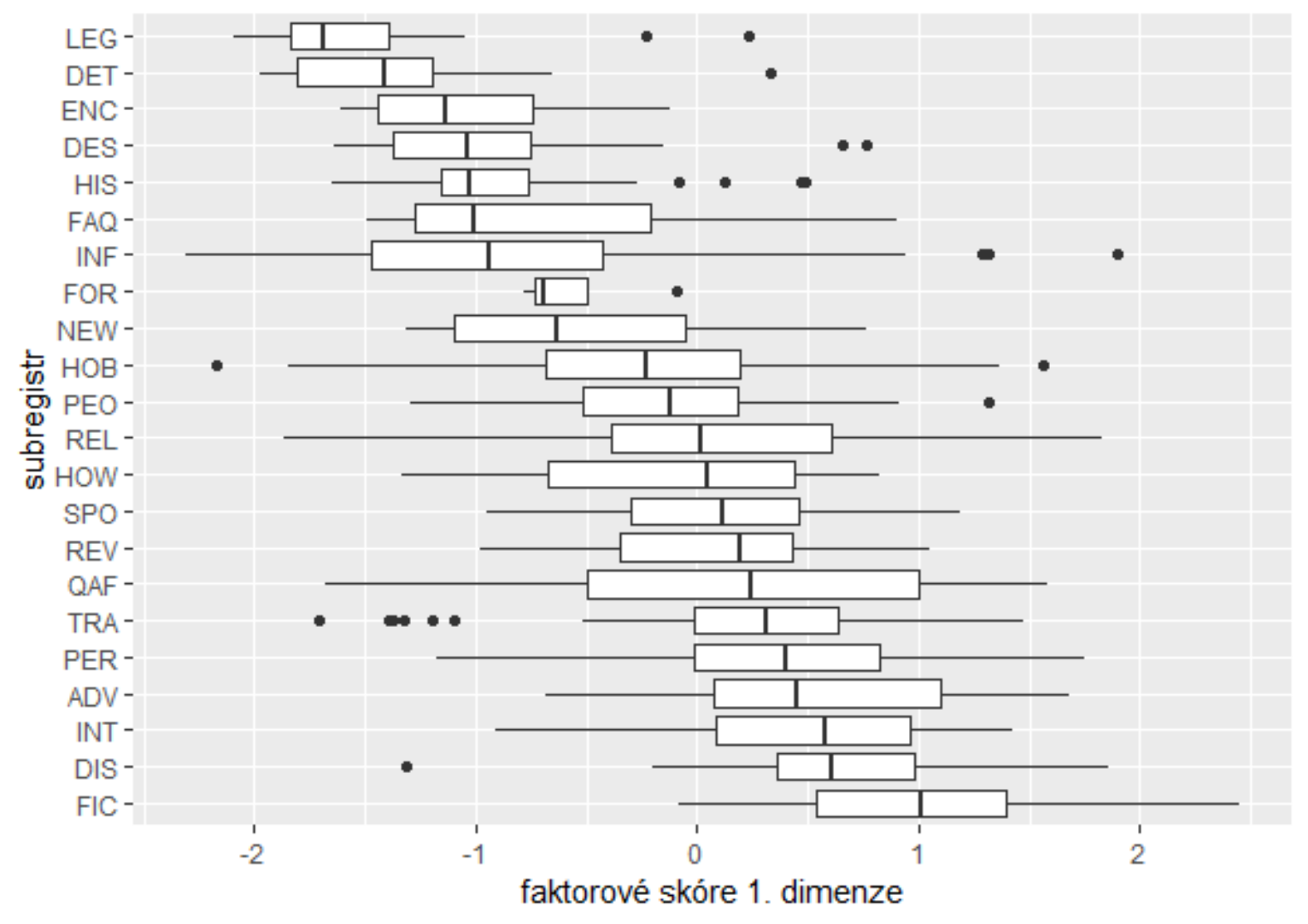
I (dyn+stat-) × II (spont+prep-) – centroidy



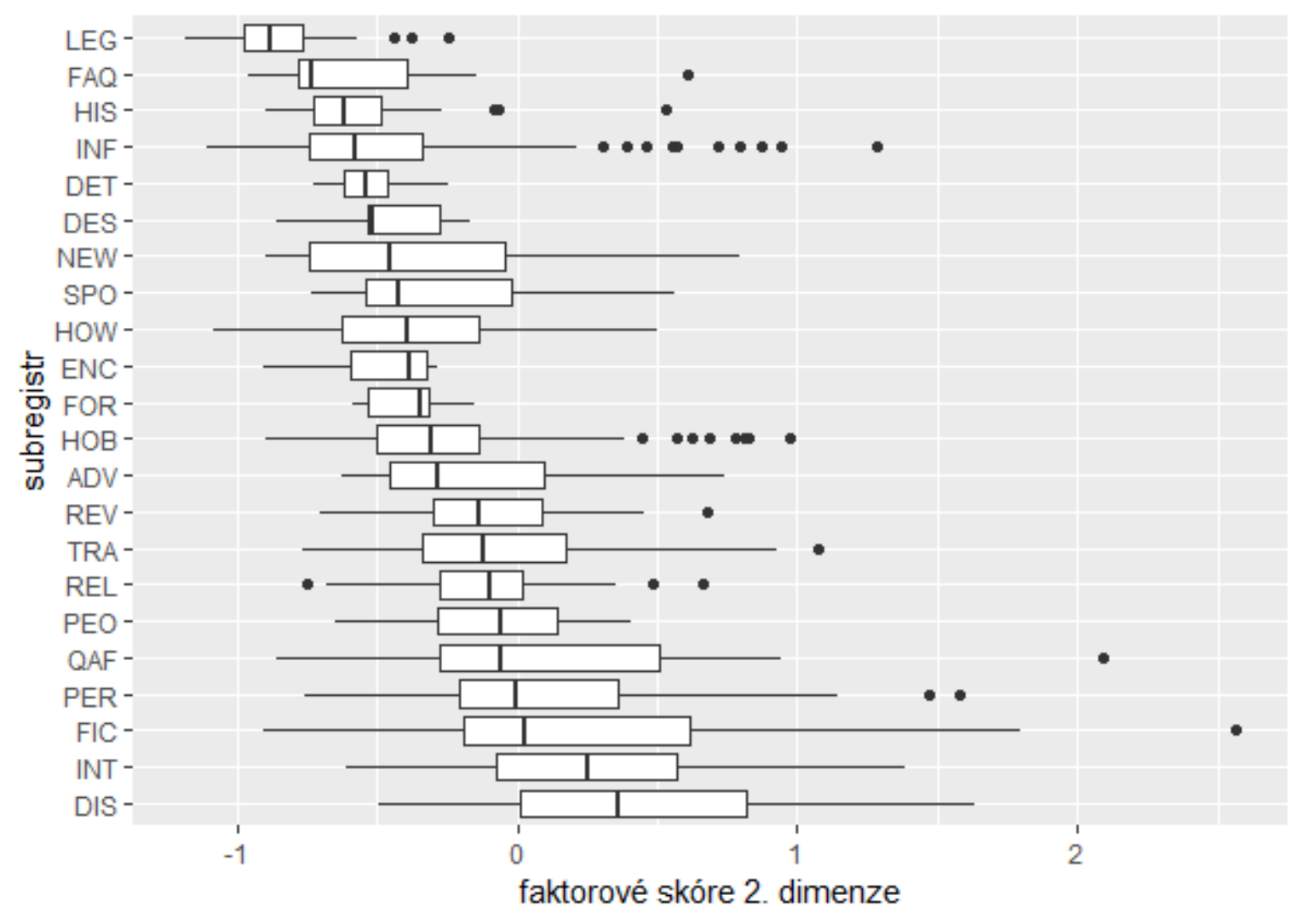
Distribuce jednotlivých subregistrů I × II



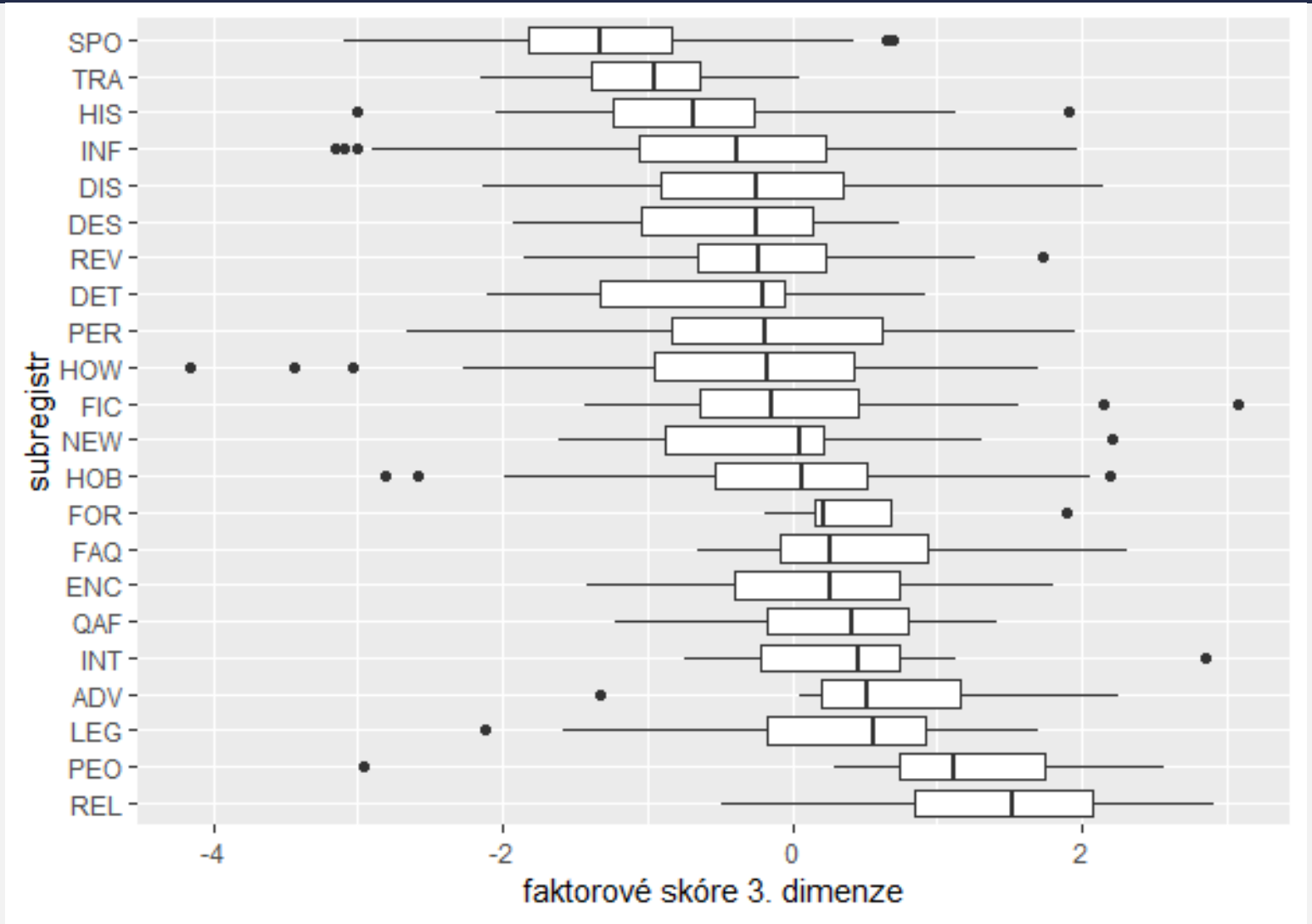
1. dynamický (+) × statický (-)



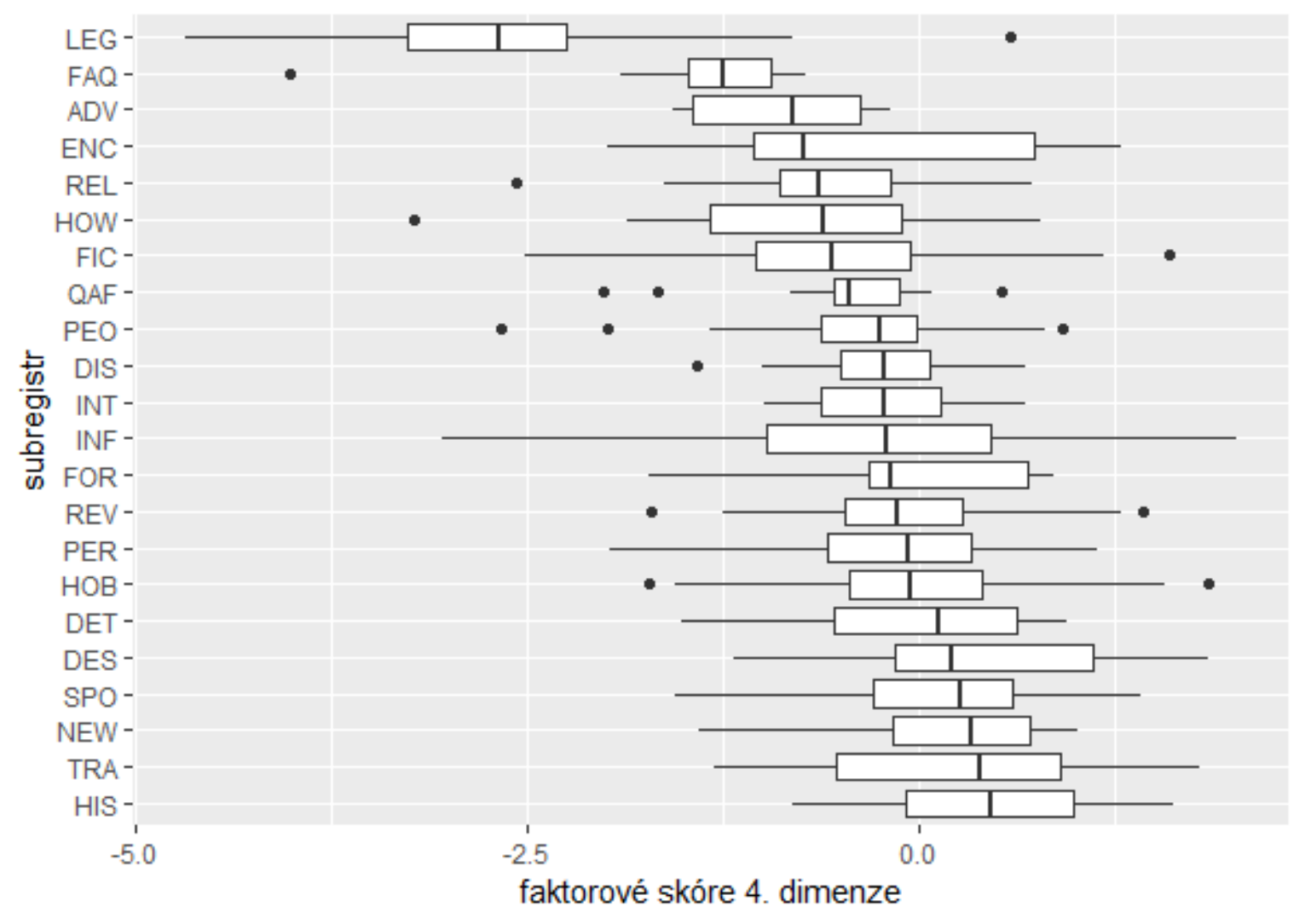
2. spontánní (+) × připravený (-)



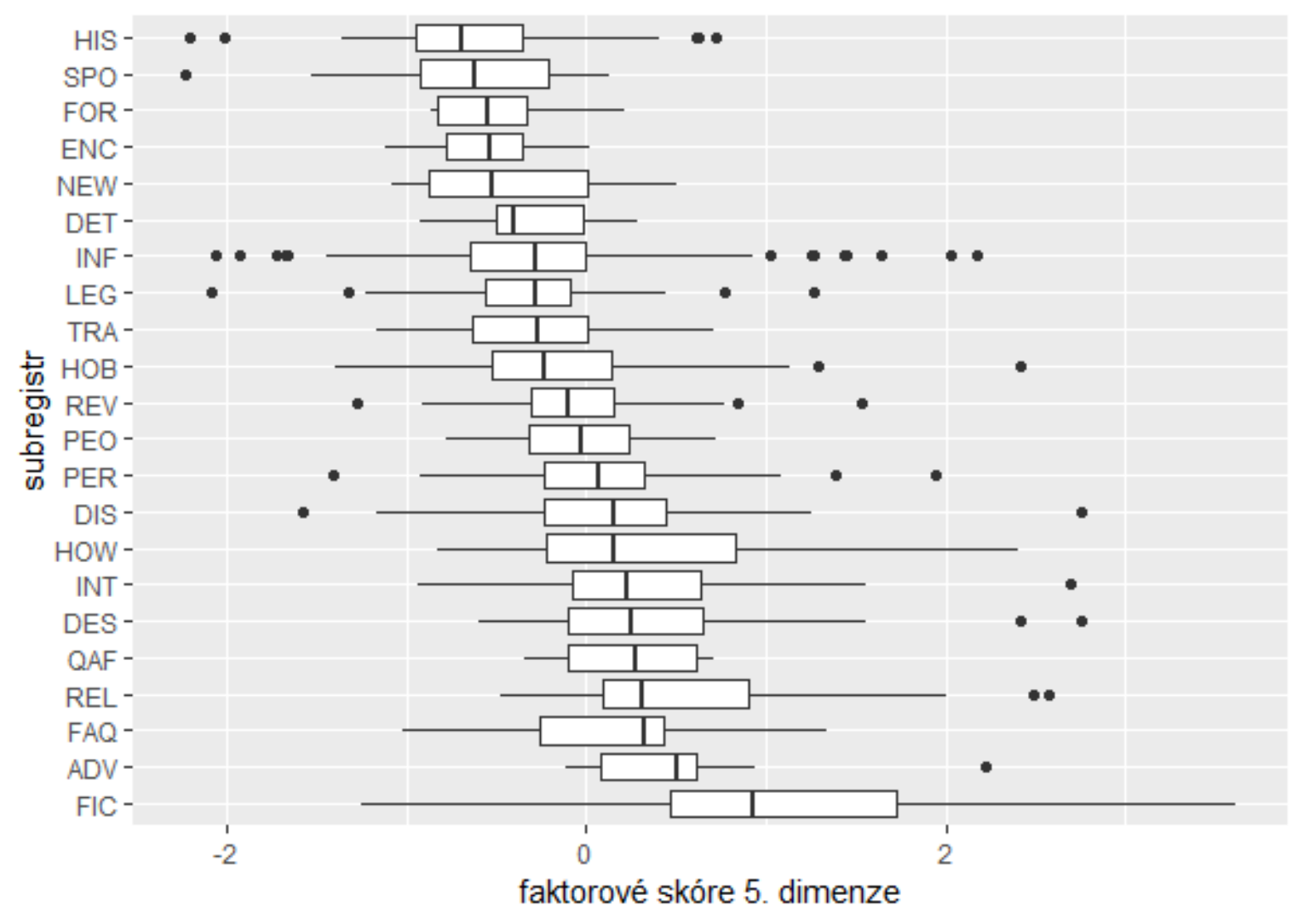
3. vyšší (+) × nižší (-) stupeň koheze



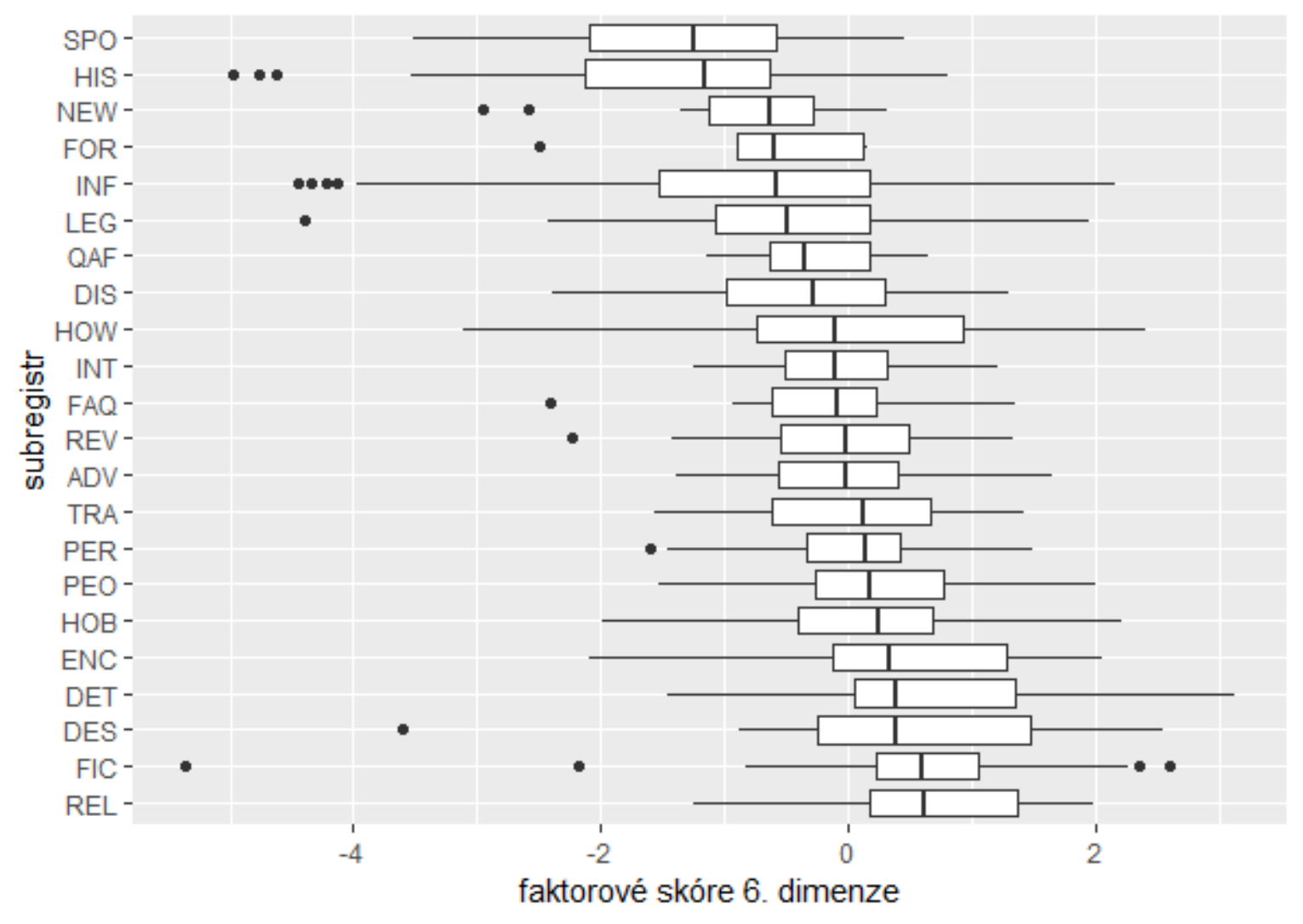
4. polytematický (+) × monotematický (-)



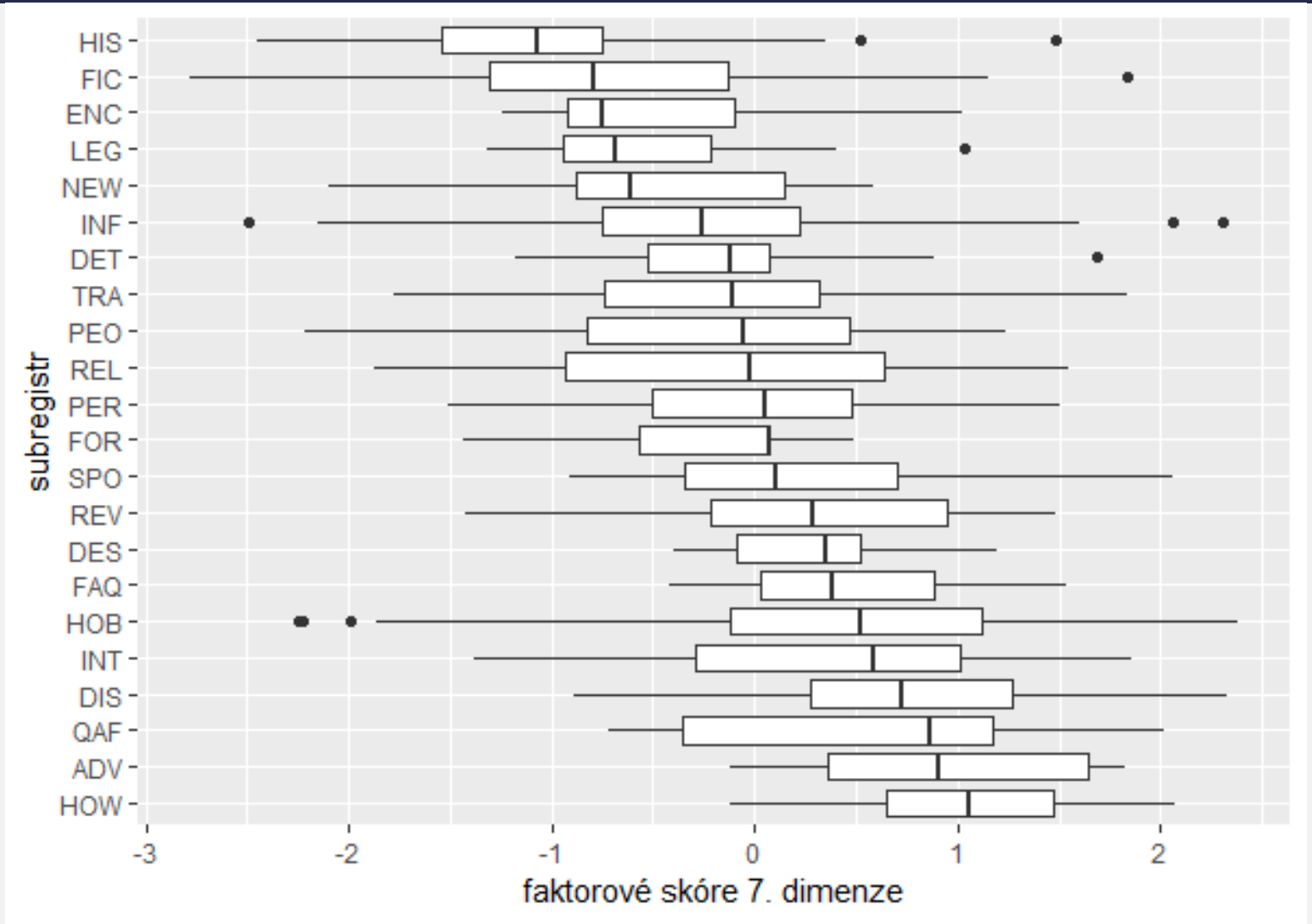
5. vyšší (+) × nižší (-) míra adresnosti



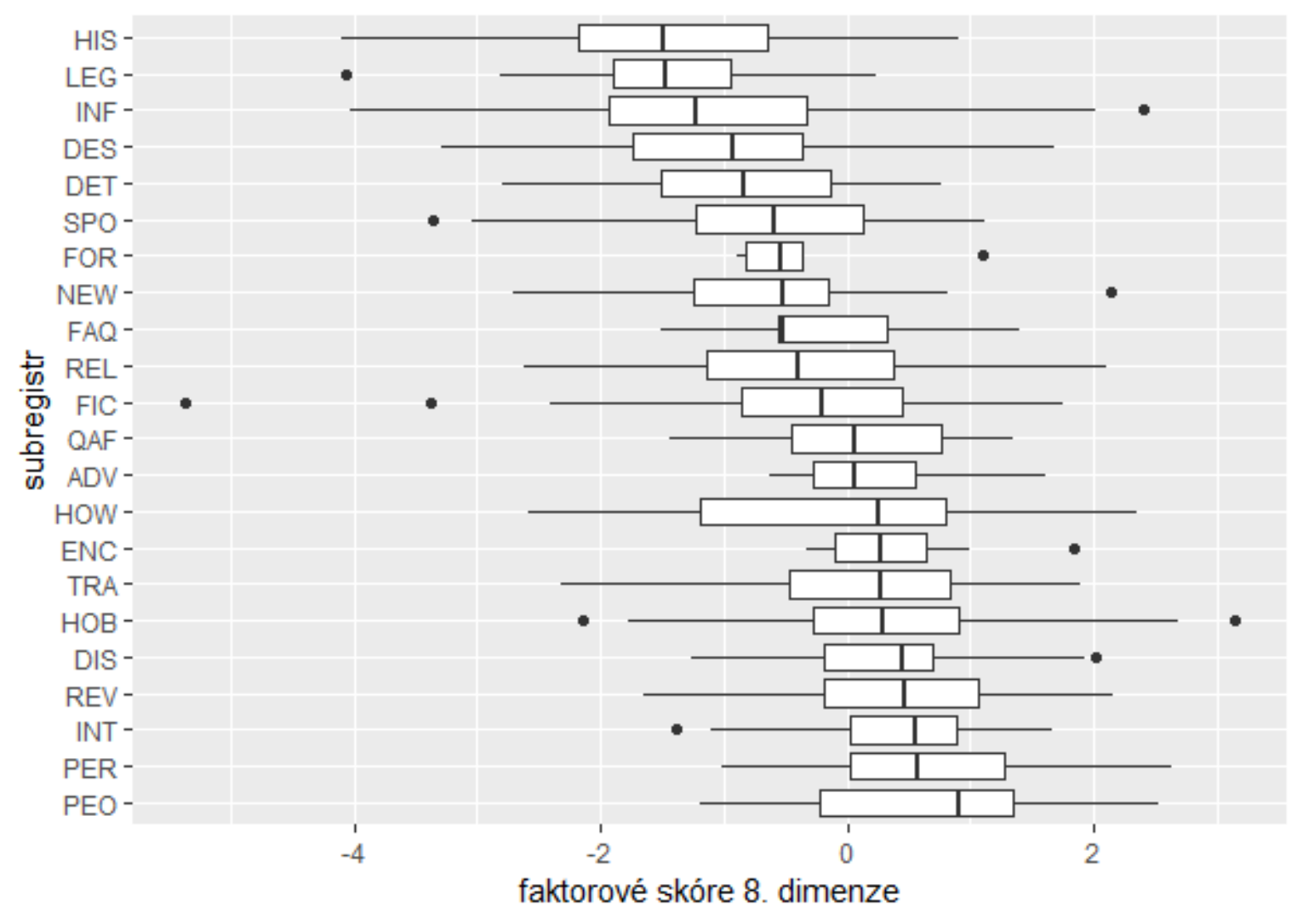
6. obecný (+) × konkrétní



7. prospektivní (+) × retrospektivní (-)



8. postojový (+) × faktuální (-)



Co s tím?

- Charakteristika subregistrů z hlediska výsledků MDA
- Možnosti vizualizace dat
- Testování vhodnosti kategorizace
  - histogramy
  - Hartigan–Hartigan dip test
  - Koeficient bimodality

Charakteristika subregistrů

Subregistr	Zkratka	Poznámka	Charakteristika textu
Poradenský text	ADV		monotematický, vyšší míra explicitní adresnosti, prospektivní
Popis s účelem prodeje	DES		statický
Popis předmětu	DET	+ popis osoby	statický
Diskusní fórum	DIS		dynamický, spontánní, prospektivní
Encyklopedický článek	ENC	+ vědecký článek	statický, retrospektivní
Nejčastěji kladené dotazy (FAQs)	FAQ		připravený, monotematický
Umělecká próza	FIC	+ písňové texty	dynamický, vyšší míra explicitní adresnosti, obecný, retrospektivní
Formální projev	FOR		nižší míra explicitní adresnosti
Historický článek	HIS		připravený, polytematický, nižší míra explicitní adresnosti, konkrétní, retrospektivní, faktuální
Hobby a volný čas	HOB		—
Návod typu jak na to, instruktáž	HOW		prospektivní
Informační text	INF		faktuální
Interview	INT		dynamický, spontánní
Právní text	LEG		statický, připravený, monotematický, faktuální
Zpravodajský článek/blog	NEW		konkrétní
Názorový blog	PEO		vyšší stupeň koheze, postojovost
Osobní blog	PER		postojovost
Fórum typu question-answer	QAF		prospektivní
Náboženský blog / kázání	REL	+ ezoterie a parapsych.	vyšší stupeň koheze, obecný
Recenze	REV		postojovost
Sportovní reportáž	SPO		nižší stupeň koheze, nižší míra explicitní adresnosti, konkrétní
Cestovatelský blog	TRA		polytematický

3D Vizualizace

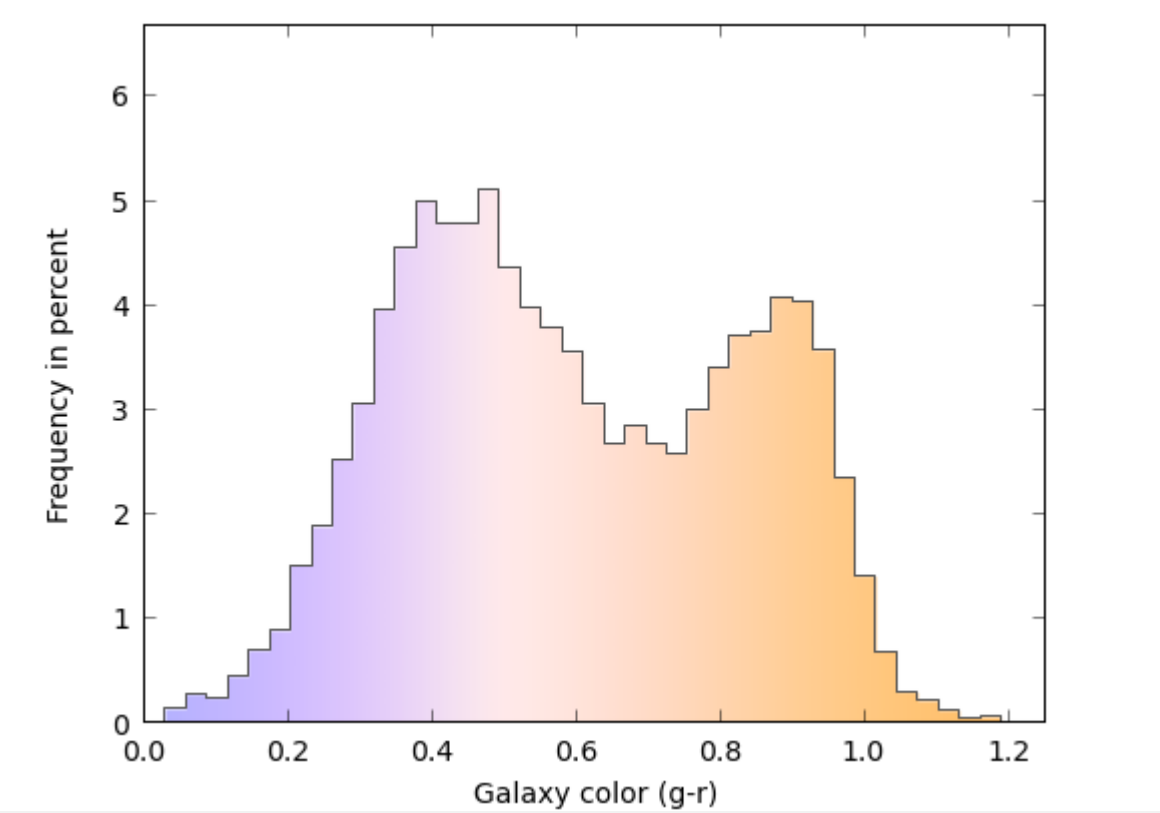
- link

Testování unimodality distribuce

- Vizualizace v podobě houslových grafů (violin plots)
- Hartigan–Hartigan dip test (Hartigans‘ dip test)
- Koeficient bimodality

Bimodalita

- Distribuce barevného spektra pozorovaných galaxií



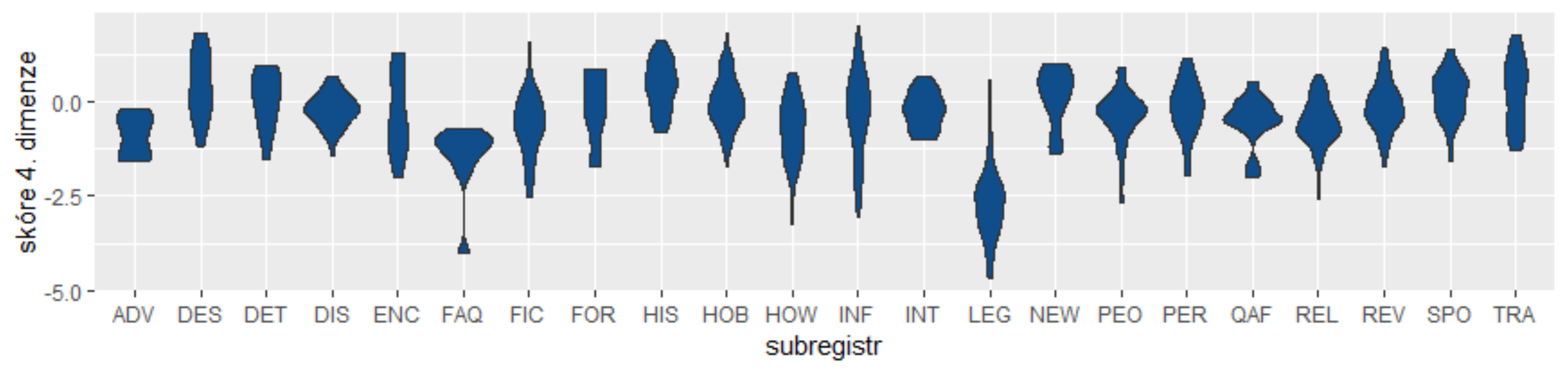
Hartigans‘ dip test (Hartigan–Hartigan 1985)

SUBREG	GLS1_D	GLS2_D	GLS3_D	GLS4_D	GLS5_D	GLS6_D	GLS7_D	GLS8_D	count
ADV	0,0846348	0,09621	0,07129	0,14209	0,13264	0,06411	0,12582	0,06373	10
DES	0,0609688	0,08159	0,05253	0,07054	0,04166	0,07492	0,06882	0,04996	21
DET	0,1226279	0,07888	0,07446	0,07276	0,09797	0,07541	0,0661	0,07511	13
DIS	0,029487	0,03494	0,04583	0,0365	0,05296	0,04801	0,02771	0,06799	44
ENC	0,0844109	0,08548	0,09898	0,10565	0,0725	0,09022	0,11577	0,08242	11
FAQ	0,0884385	0,0873	0,11068	0,10641	0,11844	0,07982	0,10211	0,07742	11
FIC	0,0379611	0,04906	0,02523	0,03279	0,03807	0,02848	0,02927	0,04086	59
FOR	0,1	0,15623	0,1	0,16749	0,11217	0,14221	0,1	0,12248	5
HIS	0,0271906	0,0392	0,04374	0,03925	0,0361	0,05213	0,03731	0,03377	55
HOB	0,0202743	0,01838	0,01688	0,02734	0,02225	0,02537	0,02741	0,02113	134
HOW	0,0560782	0,04926	0,03899	0,05895	0,06281	0,06222	0,07154	0,05505	26
INF	0,0199338	0,01228	0,01936	0,01652	0,01869	0,01738	0,01651	0,01863	244
INT	0,0553621	0,04787	0,07048	0,05771	0,0424	0,04604	0,05953	0,04743	31
LEG	0,0421018	0,04674	0,04785	0,03463	0,04474	0,0379	0,0489	0,03084	51
NEW	0,0732602	0,06564	0,10007	0,0735	0,07857	0,07087	0,07914	0,06646	15
PEO	0,0537991	0,05454	0,07281	0,03997	0,06475	0,07645	0,04723	0,05123	30
PER	0,0332566	0,04445	0,02958	0,04522	0,04047	0,03339	0,02629	0,03425	61
QAF	0,0786045	0,05593	0,06657	0,07077	0,07464	0,07549	0,10074	0,07863	13
REL	0,0604407	0,03098	0,04413	0,04883	0,0459	0,046	0,04626	0,05146	49
REV	0,0672465	0,03204	0,05128	0,03811	0,04377	0,03729	0,05295	0,03378	45
SPO	0,0460911	0,0444	0,04788	0,05599	0,06732	0,05523	0,05047	0,04451	34
TRA	0,0452194	0,03905	0,05379	0,04945	0,05045	0,04547	0,05646	0,04123	38

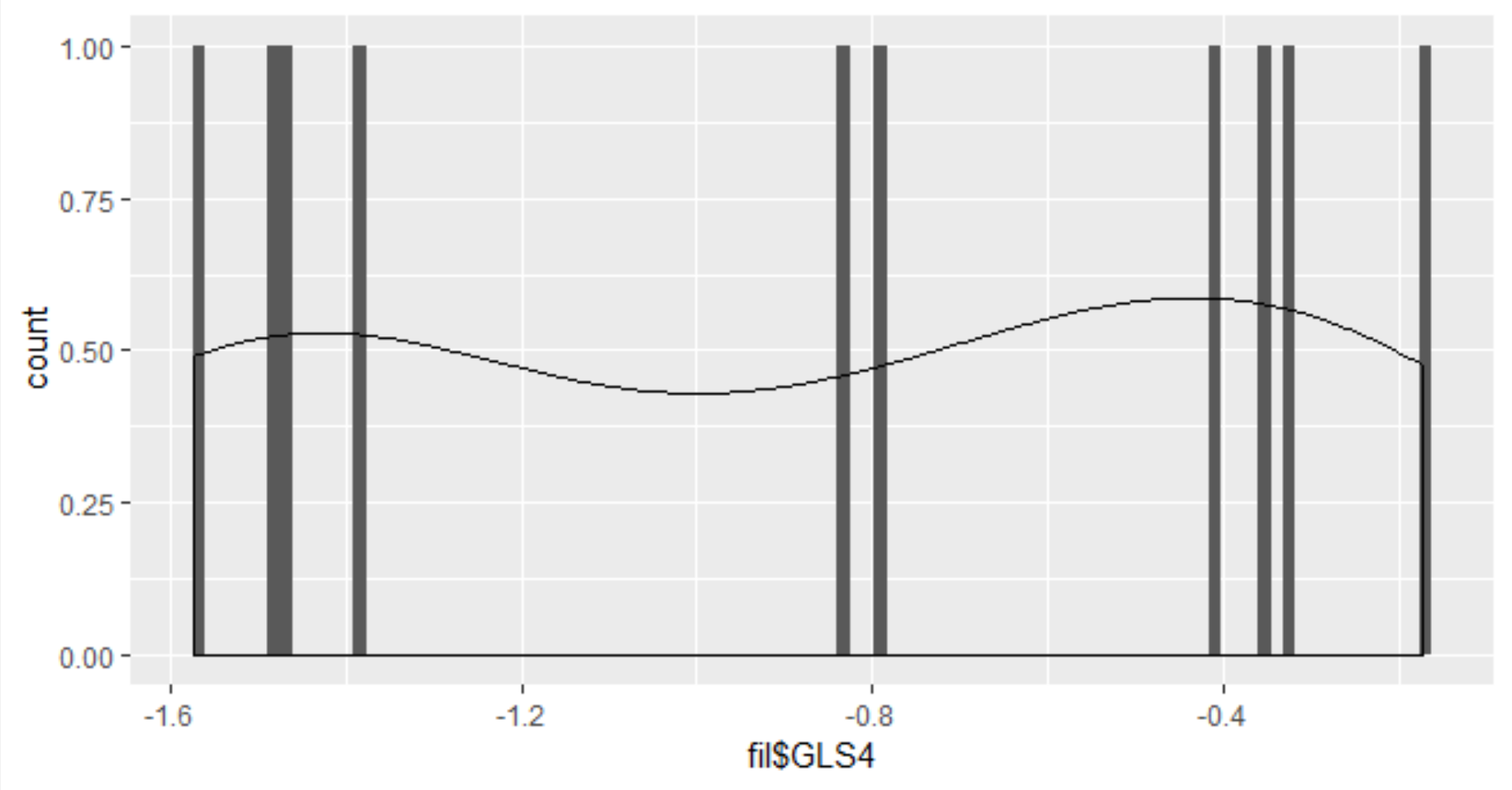
Hartigans‘ dip test – p-hodnoty

SUBREG	GLS1_P	GLS2_P	GLS3_P	GLS4_P	GLS5_P	GLS6_P	GLS7_P	GLS8_P	count
ADV	0,80887	0,54559	0,95243	0,04344	0,0881	0,98255	0,13993	0,98346	10
DES	0,79169	0,28388	0,93837	0,53362	0,99347	0,42463	0,5811	0,96398	21
DET	0,06644	0,71631	0,80125	0,83077	0,31051	0,78318	0,93147	0,78892	13
DIS	0,99362	0,97854	0,71664	0,96013	0,456	0,63627	0,99531	0,1099	44
ENC	0,73723	0,71626	0,3935	0,29972	0,9182	0,60572	0,1837	0,77635	11
FAQ	0,64855	0,67606	0,23908	0,29048	0,16002	0,82077	0,34948	0,85641	11
FIC	0,80089	0,35568	0,99442	0,9471	0,79657	0,99069	0,98858	0,68397	59
FOR	0,90286	0,23525	0,90286	0,15926	0,57312	0,3392	0,90286	0,49279	5
HIS	0,99316	0,80193	0,62139	0,80036	0,89903	0,30875	0,861	0,94954	55
HOB	0,98562	0,99254	0,99511	0,71751	0,95142	0,82713	0,71375	0,97329	134
HOW	0,77175	0,91819	0,99237	0,69157	0,57555	0,59315	0,34242	0,80017	26
INF	0,76881	0,99563	0,81127	0,95779	0,85395	0,92541	0,95811	0,85741	244
INT	0,66515	0,87191	0,24753	0,58979	0,96403	0,91114	0,53194	0,88238	31
LEG	0,74712	0,56541	0,5216	0,95435	0,64483	0,8842	0,48149	0,99033	51
NEW	0,72416	0,87222	0,19613	0,71962	0,61358	0,76903	0,60133	0,85569	15
PEO	0,73797	0,71631	0,22224	0,98798	0,4047	0,16655	0,90282	0,81026	30
PER	0,92504	0,50136	0,98286	0,47104	0,67327	0,92208	0,99274	0,9025	61
QAF	0,72169	0,9904	0,92586	0,86528	0,79795	0,78166	0,27079	0,7212	13
REL	0,1821	0,99076	0,70112	0,51755	0,63154	0,62785	0,61741	0,42176	49
REV	0,11108	0,99087	0,49592	0,92756	0,77334	0,94398	0,43811	0,98449	45
SPO	0,86712	0,90653	0,8227	0,56771	0,25286	0,59289	0,74813	0,90463	34
TRA	0,83071	0,95936	0,54811	0,7002	0,6653	0,82411	0,45874	0,92518	38

Vizualizace – violin plots (4. dimenze)



Vizualizace – histogram subregistru ADV (4. dimenze)



Koeficient bimodality (Pfister et al. 2013)

- Hodnoty testu převyšující 5/9 (BC > 0,555) indikují bimodální distribuci
- Negativní korelace s velikostí vzorku

SUBREG	GLS1	GLS2	GLS3	GLS4	GLS5	GLS6	GLS7	GLS8	count
ADV	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	10
DES	**0,681**	0,367	**0,574**	**0,578**	**0,705**	0,407	0,408	0,272	21
DET	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	13
DIS	0,296	0,493	0,411	0,344	0,251	**0,592**	0,419	0,386	44
ENC	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	11
FAQ	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	11
FIC	0,456	0,511	0,437	0,318	0,410	0,456	0,353	0,431	59
FOR	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	5
HIS	0,551	0,522	0,267	0,534	0,263	0,449	0,373	0,400	55
HOB	0,338	0,524	0,331	0,365	0,426	0,379	0,433	0,339	134
HOW	**0,759**	0,400	0,555	0,384	0,452	0,432	0,511	**0,580**	26
INF	0,444	0,506	0,367	0,454	0,266	0,395	0,335	0,373	244
INT	0,545	0,412	0,392	**0,571**	0,450	0,379	0,503	0,498	31
LEG	**0,602**	0,489	0,421	0,307	0,199	0,241	0,454	0,269	51
NEW	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	15
PEO	0,407	0,520	**0,555**	0,444	0,473	0,323	0,511	**0,568**	30
PER	0,417	0,471	0,401	0,398	0,239	0,353	0,398	0,426	61
QAF	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	13
REL	0,325	0,297	0,442	0,290	**0,583**	0,448	**0,576**	0,417	49
REV	0,528	0,415	0,352	0,335	0,285	0,371	**0,574**	0,482	45
SPO	0,423	**0,697**	0,415	0,375	0,457	0,510	0,540	0,422	34
TRA	**0,556**	0,515	0,436	**0,589**	0,418	0,527	0,446	0,438	38

BC > .555

- popis s účelem prodeje (DES)
  - BC = .685 v 1. dimenzi (statický vs. dynamický)
  - BC = .705 v 5. dimenzi (míra kódování adresáta)
- diskusní fórum (DIS)
  - BC = .592 v 6. dimenzi (obecný vs. konkrétní)
- návod typu „jak na to“, instruktáž (HOW)
  - BC = .759 v 1. dimenzi (statický vs. dynamický)

BC > .555

- Právní texty (LEG)
  - BC = .602 v 1. dimenzi (statický vs. dynamický)
- Sportovní článek/reportáž (SPO)
  - BC = .697 v 2. dimenzi (spontánní vs. připravený)

Co dál?

- Před provedením výzkumu ve větším měřítku s měřením mezianotátorské shody:
  - ověřit unimodalitu distribuce
  - ověřit překryvy blízce definovaných subregistrů

- problém s metrikami u méně zastoupených subregistrů
  - větší rozah pilotního výzkumu?

Reference

- Cvrček, V. – Komrsková, Z. – Lukeš, D. – Poukarová, P. – Řehořková, A. – Zasina, A. J. (2018b): Variabilita češtiny: multidimenzionální analýza. Slovo a slovesnost, 79(3), s. 293–321
- Henyš, J. (2019): Registrová variabilita českých internetových textů. FF UK. [diplomová práce]
- Pfister, R. et al. (2013): Good things peak in pairs: a note on the bimodality coefficient. Frontiers In Psychology 4, s. 700.
- Hartigan, J. A.–Hartigan, P. M. (1985): The Dip Test of Unimodality. The Annals of Statistics 13/1, s. 70.