Katedra obecné lingvistiky
Univerzita Palackého v Olomouci

Projekt Korpus:
Komparácia frekvencie funkčných slov v dvoch korpusoch

2023/2024

Obsah

Vypracovanie projektu Korpus

Predpokladaná časová náročnosť: 8hod
Skutočná časová náročnosť: 3hod

Úvod

Cieľom projektu Korpus je preskúmanie odlišnosti používania funkčných slov v hovorom a písanom jazyku. Aby sme prípadnú odlišnosť odhalili, budeme porovnávať frekvencie funkčných slov v dvoch rozdielnych korpusoch českého jazyka za pomoci webového rozhrania pre prácu s korpusmi KonText (nástroj je voľne dostupný na: https://www.korpus.cz/kontext). Extrahované dáta potom štatisticky spracujeme v programe Excel.

Metóda

Prv, než poskytneme návod na riešenie úlohy, musíme určiť predpoklady, s ktorými budeme pracovať.

Najdôležitejším krokom v našej práci je určenie definície funkčného slova. Funkčné slovo by z lingvistickej definície malo byť neplnovýznamové slovo plniace gramatickú funkciu. Bežne sa však v korpusoch funkčné slová špeciálne neoznačujú, a preto musíme zaviesť inú, pracovnú definíciu funkčného slova: Funkčným slovom budeme nazývať akékoľvek slovo, ktoré nespadá do žiadneho zo slovných druhov či gramatických kategórií v tabuľke 1.

	Slovný druh / gramatická kategória	CQL pos tag
1	Substantíva	N
2	Adjektíva	A
3	Adverbiá	D
4	Číslovky	C
5	Slovesá	V
6	Segmenty	S
7	Skratky	B
8	Interpunkcia	Z
9	Neznáme	X

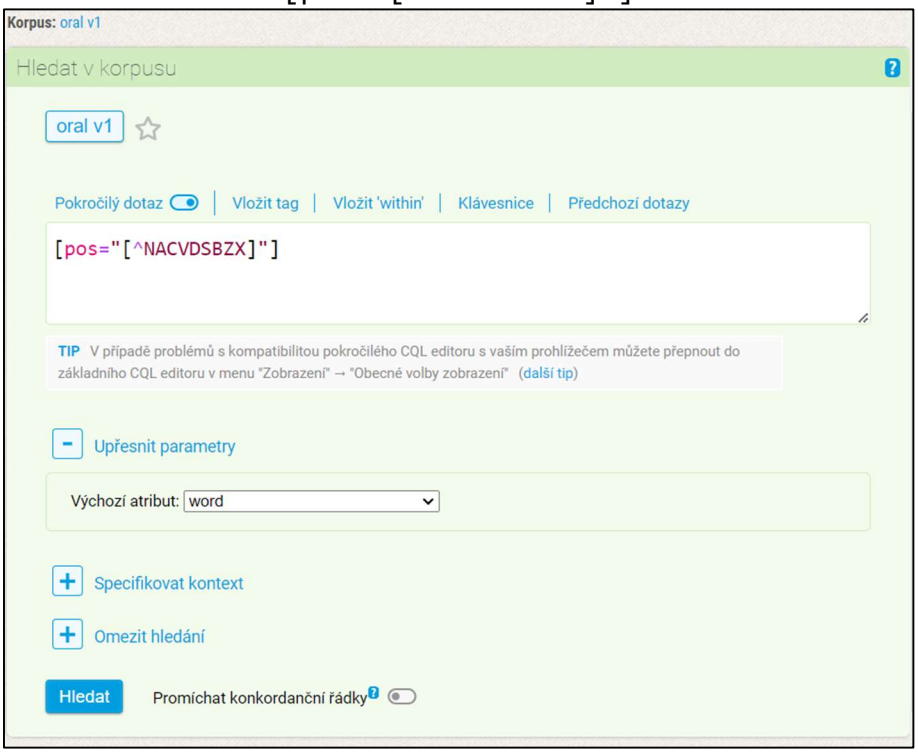
Tabuľka 1: Slovné druhy a ich CQL pos označenie v rámci korpusov ČNK (zdroj: https://wiki.korpus.cz/doku.php/seznamy:tagy).

Ďalší predpoklad, s ktorým pri práci počítame, je nižšia vývojová dynamika funkčných slov. Na vypracovanie projektu totiž použijeme dva korpusy – korpus hovorového jazyka oralv1 (informácie o veľkosti a zložení sú dostupné tu: https://wiki.korpus.cz/doku.php/cnk:oral), korpus písaného jazyka syn2020 (informácie o veľkosti a zložení sú dostupné tu: https://wiki.korpus.cz/doku.php/cnk:syn2020). Korpus oralv1 bol zostavený z nahrávok z rokov 2002–2011, korpus syn2020 je zameraný najmä na texty z rokov 2015–2019. Napriek tomu, že oba korpusy pokrývajú iné časové obdobie, veríme, že vývoj funkčných slov je natoľko pomalý, že v tak malom časovom horizonte nemohla nastať významná zmena v používaní funkčných slov. Posledným problémom by mohla byť nevyváženosť korpusu oralv1 – sú v ňom totiž obsiahnuté prevažne nahrávky stredočeských a severovýchodočeských hovoriacich. Korpus teda zväčšuje vplyv českého dialektu. Myslíme si však, že užívanie funkčných slov nie je v nijakej vyššej miere závislé od dialektu, ale vyskytuje sa naprieč celým jazykom v približne rovnakej miere. Tieto tvrdenia však vyžadujú ďalší lingvistický výskum. Náš návod je nezávislý od použitých korpusov, a teda môže byť aplikovaný aj na iné, vhodnejšie korpusy.

V našom projekte budeme pri štatistických výpočtoch pracovať s hladinou významnosti 0,05; $\alpha = 0{,}05$. Čiže tie javy, ktoré sa dejú v menej než 5 % prípadov nebudeme považovať za prejav nejakého generátora, za nenáhodné.

1. Po zaregistrovaní sa a prihlásení do služby KonText zaklikneme možnosť *Pokročilý dotaz* umožňujúcu zložitejšie dopyty pomocou dopytovacieho jazyka CQL na stránke https://www.korpus.cz/kontext/. Kliknutím na modrý názov predvoleného korpusu otvoríme ponuku korpusov a zvolíme si korpus oral v1. Do poľa pre dopyt zadáme nasledovný CQL výraz označujúci všetky funkčné slová podľa našej definície z úvodu tejto časti (viď obrázok 1); Výraz je zložený z pos tagov z tabuľky 1:

```
[pos="[^NACVDSBZX]"]
```

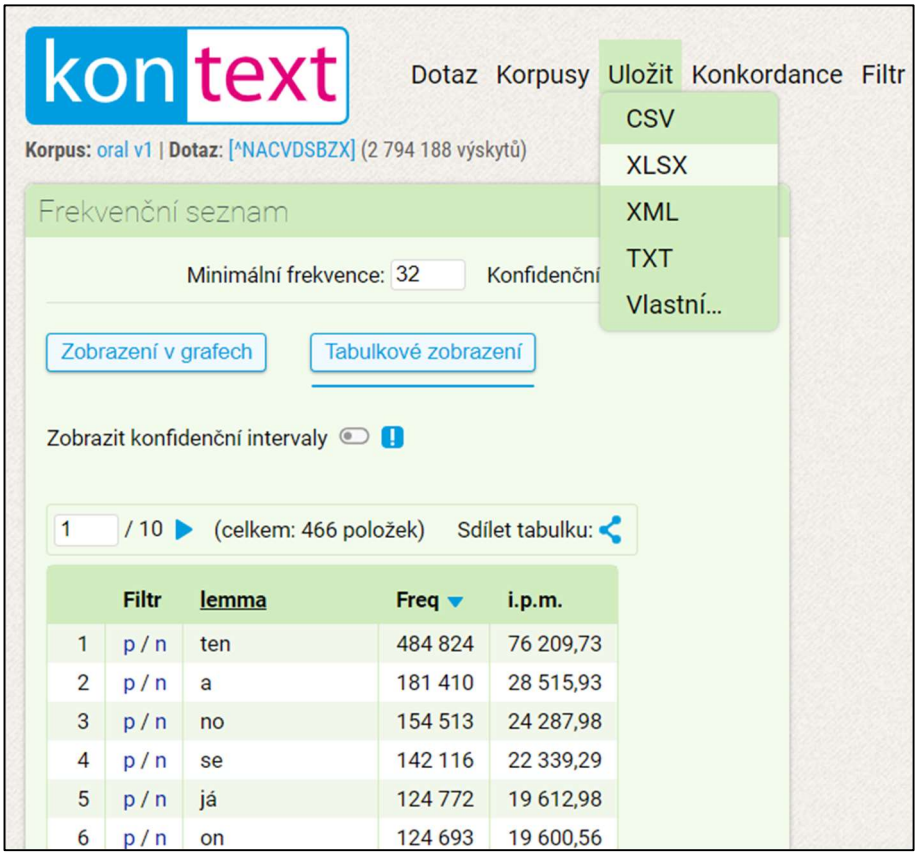


Obrázok 1: Vyplnenie úvodnej stránky webového nástroja KonText CQL dopytom.

2. Po stlačení tlačidla *Hledat* zvolíme v hornej ponuke možnosť *Frekvence*, *Lemmata* (viď obrázok 2).

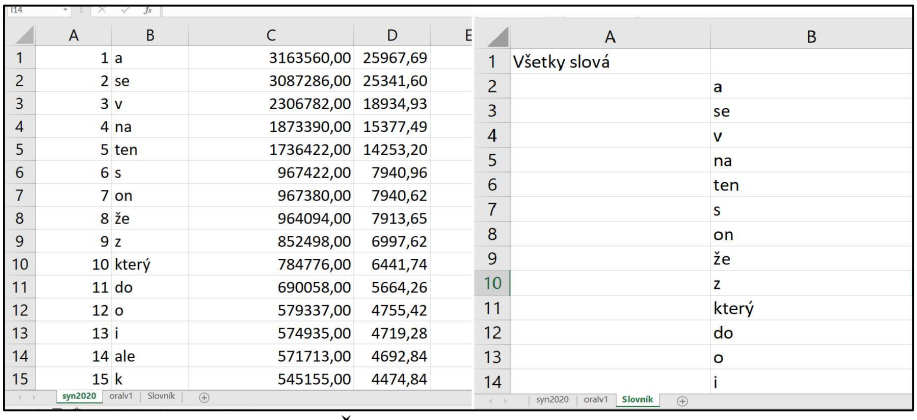
Obrázok 2: Získanie frekvencií základných, slovníkových tvarov slov (lemmata).

3. V ponuke zvolíme možnosť *Uložiť*, *XLSX* (viď obrázok 3). V našom riešení sme sa rozhodli pre zjednodušenie riešenia uvažovať len tie slová, ktoré dosahujú 5 výskytov na milión korpusových vstupov (5 i.p.m.). V prípade oralv1 sme teda stanovili minimálnu frekvenciu 32, pre korpus syn2020 je minimálna frekvencia 609.



Obrázok 3: Stiahnutie údajov o slovách.

4. Kroky 1, 2 a 3 opakujeme aj pre korpus syn2020.
5. Obsah oboch excelovských súborov nakopírujeme do dvoch rozdielnych hárkov nového excelovského súboru a pomenujeme ich podľa korpusov. Následne skopírujeme všetky slová (z oboch súborov) do tretieho hárka, v tomto hárku si postupne vytvoríme slovník funkčných slov (viď obrázok 4).



Obrázok 4: Štruktúra excelovského súboru.

6. V hárku *Slovník* vytvoríme zoznam unikátnych slov pomocou vzorca:

```
=UNIQUE(B:B),
```

pokiaľ máme slová v inom rozsahu ako celý stĺpec B, použijeme náš rozsah.

7. Vyhľadáme si frekvencie unikátnych slov v oboch korpusoch pomocou formúl:

```
=XLOOKUP(D3;'syn2020'!B:B;'syn2020'!C:C;0)
```

```
=XLOOKUP(D3; oralv1!B:B;&oralv1!C:C;0),
```

Vypočítané bunky natiahneme a aplikujeme na celý stĺpec. Tieto formuly predpokladajú, že máme slová uložené v hárkoch syn2020 a oralv1 v stĺpci B, a ich frekvenciu máme v stĺpci C. V prípade, že sa tieto slová v danom korpuse nenachádzajú Excel zapíše ako ich frekvenciu hodnotu 0.

8. Nájdeme slová, ktoré sa vyskytujú v oboch korpusoch pomocou nasledujúcej formuly:

```
=FILTER(D3:D626;F3:F626*G3:G626),
```

Rozsahy D3:D626, F3:F626, G3:G626 označujú rozsahy unikátnych slov a ich frekvencií v jednotlivých korpusoch.

9. Zopakujeme krok 7, avšak zmeníme počiatočný stĺpec z D3 na stĺpec, v ktorom máme unikátne slová nachádzajúce sa v oboch korpusoch. V našom prípade sa teda jedná o bunku I3 a jej príslušný stĺpec. Všetky kroky vytvorenia slovníku môžeme vidieť na obrázku 5.

	A	B	C	D	E	F	G	H	I	J	K
1	Všetky slová		Unikátne slová								
2		a		0		Syn2020Freq	OralV1Freq		Spoločné slova	Syn2020Fre	OralV1Freq
3		se		a		3163560	181410		a	3163560	181410
4		v		se		3087286	142116		se	3087286	142116
5		na		v		2306782	57072		v	2306782	57072
6		ten		na		1873390	75162		na	1873390	75162
7		s		ten		1736422	484824		ten	1736422	484824
8		on		s		967422	27910		s	967422	27910
9		že		on		967380	124693		on	967380	124693
10		z		že		964094	112707		že	964094	112707
11		který		z		852498	24649		z	852498	24649
12		do		který		784776	8788		který	784776	8788
13		o		do		690058	27152		do	690058	27152
14		i		o		579337	12520		o	579337	12520
15		ale		i		574935	10742		i	574935	10742
16		k		ale		571713	62406		ale	571713	62406
17		já		k		545155	9829		k	545155	9829
18		jako		já		508429	124772		já	508429	124772

Obrázok 5: Výsledná podoba Excelovského hárku slovník.

10. Teraz, keď máme informácie o frekvenciách slov vyskytujúcich sa v oboch korpusoch, môžeme vytvoriť kontingenčnú tabuľku, a vypočítať hodnotu chí-kvadrátového testu (χ^2 test). Majme označenie frekvencie v korpuse hovorového jazyka a v korpuse písaného jazyka $f_{hovorový}$, $f_{písaný}$. Potom, ak veľkosť korpusov označíme H pre hovorový a P pre písaný, dostávame doplnkové frekvencie $f^{inv}_{hovorový} = H - f_{hovorový}$, $f^{inv}_{písaný} = P - f_{písaný}$. Aby sme získali teoretickú predpokladanú relatívnu frekvenciu v jednom spojitom korpuse, sčítame veľkosti jednotlivých korpusov, a týmto súčtom vydelíme súčet frekvencií, dostávame:

$$F = \frac{f_{hovorový} + f_{písaný}}{H + P}, F^{inv} = 1 - F$$

Môžem ešte podotknúť, že hodnota $F * Veľkost\ Korpusu$ sa rovná očakávanej frekvencii slova v korpuse.

11. Hodnota chí-kvadrát testu je potom $\chi^2 = \sum_{i=1}^{4} \chi_i^2$, kde pre jednotlivé χ_i^2 platí:

$$\chi_1^2 = \frac{(f_{hovorový} - H * F)^2}{H * F}$$

$$\chi_2^2 = \frac{(f_{hovor\ \ ý}^{inv} - H * F^{inv})^2}{H * F^{inv}}$$

$$\chi_3^2 = \frac{(f_{písaný} - P * F)^2}{P * F}$$

$$\chi_4^2 = \frac{(f_{písaný}^{inv} - P * F^{inv})^2}{P * F^{inv}}$$

12. V Exceli daný výpočet vykonáme pomocou sady vzorcov. Najprv si však vytvoríme nový hárok s názvom *Štatistika*, do ktorého prekopírujeme spoločné slová oboch korpusov s ich frekvenciami. Taktiež si uložíme veľkosti korpusov. Inverzné, doplnkové frekvencie spočítame pomocou vzorcov:

`=$B$1-B4` a `=$B$2-C4`,

Kde bunky uzavreté symbolom $ pevne označujú veľkosti korpusov. Vzorec aplikujeme na celý stĺpec. Výsledok môžeme vidieť na obrázku 6.

	A	B	C	D	E
1	Veľkosť Syn2020	121826797			
2	Veľkosť Oralv1	6361707			
3	lemma	Syn2020Freq	OralV1Freq	SynInv	OralInv
4	a	3163560	181410	118663237	=$B$2-C4
5	se	3087286	142116	118739511	6219591
6	v	2306782	57072	119520015	6304635
7	na	1873390	75162	119953407	6286545
8	ten	1736422	484824	120090375	5876883
9	s	967422	27910	120859375	6333797
10	on	967380	124693	120859417	6237014
11	že	964094	112707	120862703	6249000
12	z	852498	24649	120974299	6337058
13	který	784776	8788	121042021	6352919
14	do	690058	27152	121136739	6334555
15	o	579337	12520	121247460	6349187
16	i	574935	10742	121251862	6350965
17	ale	571713	62406	121255084	6299301
18	k	545155	9829	121281642	6351878
19	já	508429	124772	121318368	6236935
20	jako	427362	87057	121399435	6274650
21	co	405032	38687	121421765	6323020
22	pro	390840	6065	121435957	6355642
23	svůj	380451	2971	121446346	6358736
24	za	379470	15861	121447327	6345846
25	po	333165	7498	121493632	6354209
26	když	307202	30971	121519595	6330736
27	všechen	293490	11993	121533307	6349714
28	tento	292424	2835	121534373	6358872
29	od	257296	8885	121569501	6352822

syn2020 | oralv1 | Slovník | **Štatistika**

Obrázok 6: Výpočet doplnkových frekvencií.

13. Nasledujúcimi vzorcami určíme hodnoty F a F^{inv}:

```
=(B4+C4)/($B$1+$B$2),
```

kde B4 a C4 odkazujú k frekvenciám v korpusoch a $B$1 a $B$2 k veľkostiam korpusov. Aplikujeme na celý stĺpec. F^{inv} dostaneme ako doplnok ku jednotke:

```
=1-F4,
```

Kde F4 odkazuje na prvú bunku s výsledkom F. Aplikujeme na celý stĺpec.

14. Dopočítame hodnoty jednotlivých čiastkových χ_i^2 hodnôt pomocou nasledujúcich vzorcov:

```
=POWER(C4-F4*$B$2;2)/(F4*$B$2),
=POWER(E4-G4*$B$2;2)/(G4*$B$2),
=POWER(B4-F4*$B$1;2)/(F4*$B$1),
=POWER(D4-G4*$B$1;2)/(G4*$B$1),
```

Kde C4 označuje bunku s frekvenciou v korpuse oralv1, F4 označuje bunku s hodnotou F, B2 veľkosť korpusu oralv1, E4 označuje inverznú frekvenciu v korpuse oralv1, G4 označuje F^{inv}, B4 a B1 označujú frekvenciu v korpuse syn2020 a veľkosť tohto korpusu, a D4 označuje doplnkovú frekvenciu v tomto korpuse. Následne ešte aplikujeme vzorec:

```
=SUM(H4:K4),
```

Kde bunky H4 až K4 označujú jednotlivé čiastkové hodnoty χ_i^2. P-hodnotu získame aplikovaním excelovskej funkcie CHISQ.DIST.RT:

```
=CHISQ.DIST.RT(L4;1),
```

Kde L4 označuje bunku s výslednou sumou χ_i^2, a 1 označuje počet stupňov voľnosti. Všetky vzorce aplikujeme na celé stĺpce. Ukážku výslednej podoby súboru môžeme vidieť na obrázku 7.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Veľkosť Syn2020	121826797											
2	Veľkosť Oralv1	6361707											
3	lemma	Syn2020Freq	OralV1Freq	SynInv	OralInv	F	Finv	Chi1	Chi2	Chi3	Chi4	Chi-Test	p-value
4	a	3163560	181410	118663237	6180297	0,02609415	0,97390585	1429,883	38,31128	74,66746	2,000587385	1544,862	0
5	se	3087286	142116	118739511	6219591	0,025192602	0,974807398	2055,891	53,13177	107,3571	2,774502622	2219,154	0
6	v	2306782	57072	119520015	6304635	0,018440452	0,981559548	30933,96	581,153	1615,349	30,34738901	33160,81	0
7	na	1873390	75162	119953407	6286545	0,015200677	0,984799323	4798,052	74,05939	250,5508	3,867327528	5126,529	0
8	ten	1736422	484824	120090375	5876883	0,017327966	0,982672034	1272881	22445,37	66468,92	1172,080917	1362967	0
9	s	967422	27910	120859375	6333797	0,007764596	0,992235404	9345,921	73,13517	488,0372	3,819065724	9910,913	0
10	on	967380	124693	120859417	6237014	0,008519274	0,991480726	91696,16	787,897	4788,307	41,1434117	97313,51	0
11	že	964094	112707	120862703	6249000	0,008400137	0,991599863	65732,09	556,836	3432,482	29,0775743	69750,48	0
12	z	852498	24649	120974299	6337058	0,006842634	0,993157366	8190,139	56,42824	427,6831	2,946641875	8677,197	0
13	který	784776	8788	121042021	6352919	0,006190602	0,993809398	23767,78	148,0534	1241,136	7,731240569	25164,7	0
14	do	690058	27152	121136739	6334555	0,005594963	0,994405037	2002,028	11,2643	104,5445	0,58821352	2118,425	0
15	o	579337	12520	121247460	6349187	0,004617083	0,995382917	9669,163	44,85041	504,9167	2,342055947	10221,27	0
16	i	574935	10742	121251862	6350965	0,004568873	0,995431127	11551,8	53,02097	603,2269	2,768716831	12210,82	0
17	ale	571713	62406	121255084	6299301	0,00494677	0,99505323	30411,36	151,1859	1588,059	7,894815993	32158,5	0
18	k	545155	9829	121281642	6351878	0,004329437	0,995670563	11392,24	49,53643	594,8943	2,586756328	12039,25	0
19	já	508429	124772	121318368	6236935	0,004939608	0,995060392	277294,1	1376,524	14480,1	71,88108079	293222,6	0
20	jako	427362	87057	121399435	6274650	0,004012989	0,995987011	148285,1	597,4641	7743,341	31,19914111	156657,1	0
21	co	405032	38687	121421765	6323020	0,003461457	0,996538543	12613,68	43,81337	658,6773	2,287902583	13318,46	0
22	pro	390840	6065	121435957	6355642	0,00309626	0,99690374	9434,958	29,30382	492,6867	1,530224286	9958,479	0
23	svůj	380451	2971	121446346	6358736	0,002991079	0,997008921	13550,25	40,65146	707,5842	2,122789922	14300,61	0
24	za	379470	15861	121447327	6345846	0,003083982	0,996916018	719,9756	2,22726	37,5966	0,116305918	759,9157	2,8E-167
25	po	333165	7498	121493632	6354209	0,002657516	0,997342484	5235,718	13,95108	273,4054	0,728515319	5523,803	0
26	když	307202	30971	121519595	6330736	0,002638091	0,997361909	11994,81	31,72709	626,3601	1,656765877	12654,55	0
27	všechen	293490	11993	121533307	6349714	0,002383076	0,997616924	661,7645	1,580803	34,55686	0,082548366	697,9847	8,2E-154

Obrázok 7: Výsledná podoba Excelovského súboru s hodnotou chí-kvadrátovej štatistiky.

15. Nakoniec označíme celú tabuľku (na obrázku 7 všetky bunky tabuľky začínajúce od A3 nadol a napravo), z ponuky nástrojov vyberieme *Vložiť*, *Tabuľka* a vytvoríme tabuľku.

Potom klikneme na stĺpec s hodnotou chí-testu a zvolíme možnosť zoradiť od najväčšieho po najmenšie. Získavame zoradené funkčné slová podľa miery ich odlišného používania. V prípade, že je p-hodnota pri funkčnom slove menšia než hladina významnosti (0,05), konštatujeme, že môžeme zamietnuť nulovú hypotézu H_0: $f_{písaný} = f_{hovor\ \ ý}$, a teda platí H_a: $f_{písaný} \neq f_{hovorový}$. V prípade, že je p-hodnota vyššia túto hypotézu zamietnuť nemôžeme. V tabuľke 2 uvádzame niekoľko zoradených funkčných slov, pri ktorých môžeme nulovú hypotézu zamietnuť, a teda sa používajú v hovorovom jazyku ináč ako v písanom jazyku. V tabuľke 3 zas uvádzame niekoľko funkčných slov, pre ktoré sa nulovú hypotézu nepodarilo zamietnuť.

	Lemma	Chí-kvadrát	p-hodnota		Lemma	Chí-kvadrát	p-hodnota
1	No	2491021,83	≈ 0	**34**	O	10221,27	≈ 0
2	Jo	1546682,85	≈ 0	**35**	Tento	10052,43	≈ 0
3	Ten	1362967,29	≈ 0	**36**	Něco	9997,95	≈ 0
4	Prostě	316258,47	≈ 0	**37**	Pro	9958,48	≈ 0
5	Já	293222,64	≈ 0	**38**	My	9953,07	≈ 0
6	Ne	201316,47	≈ 0	**39**	S	9910,91	≈ 0
7	Jako	156657,12	≈ 0	**40**	Z	8677,20	≈ 0
8	Fakt	115273,07	≈ 0	**41**	Vůbec	8366,13	≈ 0
9	On	97313,51	≈ 0	**42**	Tamten	7665,26	≈ 0
10	Viď	95137,46	≈ 0	**43**	Jenž	7279,81	≈ 0
11	Takový	81189,41	≈ 0	**44**	Při	6205,70	≈ 0
12	Nějaký	74401,38	≈ 0	**45**	Po	5523,80	≈ 0
13	Hele	71349,05	≈ 0	**46**	Jeho	5491,36	≈ 0
14	Že	69750,48	≈ 0	**47**	Hej	5134,94	≈ 0
15	Třeba	51187,29	≈ 0	**48**	Na	5126,53	≈ 0
16	Asi	50529,57	≈ 0	**49**	Či	4984,24	≈ 0
17	Aha	48993,31	≈ 0	**50**	Anebo	4448,13	≈ 0
18	Takžez	46262,58	≈ 0	**51**	Tedy	3699,86	≈ 0
19	Vždyť	33179,32	≈ 0	**52**	Mezi	3690,31	≈ 0
20	V	33160,81	≈ 0	**53**	Prý	3277,57	≈ 0
21	Ale	32158,50	≈ 0	**54**	Tenhle	3149,02	≈ 0
22	Protože	27737,01	≈ 0	**55**	Cože	2926,75	≈ 0
23	Který	25164,70	≈ 0	**56**	Tak	2770,84	≈ 0
24	Jestli	21452,63	≈ 0	**57**	Pokud	2702,67	≈ 0
25	Takovýhle	18327,88	≈ 0	**58**	Před	2634,08	≈ 0
26	Ty	17231,35	≈ 0	**59**	Nic	2353,16	≈ 0
27	Jen	14954,80	≈ 0	**60**	Se	2219,15	≈ 0
28	Svůj	14300,61	≈ 0	**61**	Podle	2166,55	≈ 0
29	Co	13318,46	≈ 0	**62**	Však	2121,51	≈ 0
30	Když	12654,55	≈ 0	**63**	Do	2118,43	≈ 0
31	Nebo	12228,22	≈ 0	**64**	Než	1950,04	≈ 0
32	I	12210,82	≈ 0	**65**	Bez	1883,10	≈ 0
33	K	12039,25	≈ 0	**66**	náš	1866,13	≈ 0

Tabuľka 2: Výsledky analýzy – funkčné slová s odlišným využitím zoradené podľa hodnoty chí-kvadrát testu pri hladine významnosti $\alpha = 5\ \%$.

	Lemma	Chí-kvadrát	p-hodnota
1	Ani	0,29	0,59
2	Nikdo	0,04	0,83
3	Jenže	0,69	0,41
4	Tvůj	0,24	0,63
5	Cosi	3,84	0,05
6	Ha	1,59	0,21
7	Něčí	1,07	0,30
8	Ó	≈0,00	0,99

Tabuľka 3: Výsledky analýzy – funkčné slová s nepreukázaným odlišným využitím v hovorom a písanom jazyku. Hladina významnosti $\alpha = 5\ \%$.

Záver

Zadaná úloha je úplne riešiteľná. Konečná podoba riešenia však závisí od definície funkčných slov a dostupných korpusov hovorového a písaného jazyka. V našom prípade sme za funkčné slová považovali všetky slová, ktorých slovný druh či kategória sa nevyskytujú v tabuľke 1. Ako korpus hovorového slova sme použili voľne dostupný korpus ČNK oralv1, ako zdroj dát písaného jazyka sme použili korpus syn2020. Pomocou komparácie frekvencií jednotlivých funkčných slov oboch korpusoch sa nám podarilo ukázať, že v prípade niektorých funkčných slov existuje štatisticky významný rozdiel v ich používaní v hovorovom a písanom jazyku.[1]

Napriek tomu, že nebolo cieľom tohto projektu aj tento rozdiel vysvetliť, môžeme postulovať niekoľko hypotéz, ktoré je treba overiť pri ďalšom bádaní. Zdá sa, že v hovorovom jazyku je vo väčšej miere využívané deiktické lexikum, a teda slová, ktoré odkazujú k momentálnemu kontextu (napr. zámená). Taktiež môžeme pozorovať väčší výskyt slov slúžiacich na upútanie pozornosti či uvedenie vety (sem patria slová ako: *Hele*, *hej*, ale aj napríklad *No*. Jedná sa o fatickú a emotívnu funkciu jazyka v Jakobsonovej teórii). Ďalší dôvod, prečo existuje rozdiel, medzi používaním niektorých funkčných slov v hovorovom a písanom jazyku, je snaha neopakovať slová v písanom jazyku a nehromadiť rovnaké slovné formy – v písanom jazyku radšej nahradíme niekoľkonásobné spojenie slov pomocou spojky *a* čiarkami a jediným *a*; V hovorovom jazyku máme väčšiu tendenciu nehľadieť na celkovú formu prejavu. Každá z týchto téz však musí byť podrobená ďalšiemu skúmaniu s využitím korpusu.

[1] Všetky dáta, aj s postupom sú voľne dostupné vo formáte Excel súboru počas doby 7 dní pomocou odkazu https://www.transfernow.net/dl/korpus_projekt_lia. Súbor je možné rozbaliť až po zadaní hesla: *korpuslingvo*. Spolu s Excel súborom je dostupný aj krátky script v jazyku R, ktorý vykonáva totožnú úlohu.