

Analýza statistických dat při korpusovém výzkumu víceslovných předložkových jednotek s místním významem (na materiálu korpusu SYN2010)

Aksana Mikalayenka
Filozofická fakulta Univerzity Karlovy v Praze
<kust_alive@mail.ru>

Abstract:

Analysis of the statistical data obtained through the corpus research of multi-word prepositional units with spatial relations (based on examples from Corpus SYN2010)

The article deals with analysis of the statistical data obtained through the corpus research of multi-word prepositional units requiring thorough study. These multi-word units consist of spatial adverbs and simple prepositions and are used in speech as spatial prepositions (for example, such units as *blízko k*, *blízko u*, *daleko od*, *nedaleko od*, *vysoko nad*). By means of excerpting from Czech grammar guides a list of spatial adverbs has been compiled. These spatial adverbs can be components of multi-word prepositional units. At the present moment this list includes more than a hundred adverbs. By using the search system of Corpus SYN2010 it has been found out if these spatial adverbs are used or not with simple prepositions the given adverbs are used with and we can answer the question *Are these combinations coincidental or fixed?* So a number of the specific spatial prepositional units *adverb + preposition* has been found. At the present moment this number is equal to 80 units. Moreover, the information about absolute and relative frequency of the given units has been obtained. The given units have been analysed with relation to their quantity. Finally, the conclusion on the role of quantitative analysis in the research of these prepositional units has been drawn.

Klíčová slova / key words:

Český národní korpus, lingvistická statistika, víceslovné předložky
Czech national corpus, multi-word prepositions, statistical linguistics

Tento příspěvek je inspirován zkušenostmi s využitím Českého národního korpusu v lingvistickém výzkumu, který autorka provádí v rámci své doktorské disertační práce. Jedná se o konfrontační výzkum místních víceslovných předložek typu *adverbium + prepozice* (dále se uvádí jako *Adv + Prep*) v ruštině, běloruštině a češtině. Zkoumají se jednotky jako například rus. *близко от* кого/чего, *высоко над* кем/чем, *где-то в* ком/чем; běl. *блізка каля* каго/чаго, *высока над* кім/чым, *недзе ў* кім/чым; čes. *blízko u* koho/čeho, *vysoko nad* kým/čím, *někde v* kom/čem aj. ČNK je v tomto výzkumu využíván jako zdroj pro zkoumání české části materiálu.

V tomto příspěvku se budeme věnovat spíše metodologickým otázkám, a to následujícím: jak je možné využít statistické funkce korpusu ve vlastním lingvistickém výzkumu a jakým způsobem mohou získaná statistická data k tomuto výzkumu přispět. Domníváme se, že na otázku „Jaká je relace mezi kvantitativní a kvalitativní stránkou objektu výzkumu?“ dříve nebo později narazí každý badatel. Ani my nejsme výjimkou a dále se pokoušíme tuto relaci vůči objektu našeho výzkumu vysledovat.

Je důležité podotknout, že české víceslovné předložky typu *Adv + Prep* nejsou v dosavadní odborné literatuře téměř popsány. V nejvýznamnějších pracích

věnovaných českým předložkám (Kroupová, 1982; Blatná, 2006) je uvedeno jen něco málo přes deset jednotek daného typu. Mezi nimi se ojediněle objevují jednotky s převážně místním významem, které jsou středem našeho zájmu. Jsou to například jednotky *stranou od* (Kroupová, 1982; Blatná, 2006), *napravo od*, *nalevo od* (Blatná, 2006). Jednotka *stranou od* je zároveň jediným útvarem typu *Adv + Prep* uvedeným mezi místními předložkami v MČ2 (Komárek et al., 1986). Daná jednotka je také zaevidována ve *Slovníku české frazeologie a idiomatiky: Výrazy neslovesné* (Čermák et al., 1988). Jednotky *stranou od* a také *daleko od* se dále objevují v ilustračním soupisu místních předložek uvedeném v PMČ (Karlík et al., 1995, s. 347). Jsou to však ojedinělé případy. Zato v ruštině a běloruštině jsou popsány desítky místních předložek daného typu (viz např. Vsevolodova et al., v tisku; Šuba, 1993; Kanjuškevič, 2008). Na tomto pozadí vzniká otázka: Proč je českých jednotek popsáno tak málo? Buď se v jazyce nevyskytují, anebo jednoduše nebyly zatím důkladně zkoumány. Pro odpověď na tuto otázku jsme využili korpus SYN2010 a pokusili jsme se pomocí jeho vyhledávacího systému a statistických funkcí stanovit alespoň přibližný soubor vyskytujících se českých místních víceslovných předložek typu *Adv + Prep*. Vycházeli jsme přitom z předpokladů, že je soubor daných jednotek v češtině mnohem větší, než je doposud popsáno, a že právě korpus a jeho statistické funkce jsou účinným nástrojem ke stanovení tohoto souboru. Výsledky této korpusové sondy budou předloženy níže.

V první řadě byl způsobem excerptce z dosavadních českých příruček a mluvnic (Komárek et al., 1986; Karlík et al., 1995; Cvrček et al., 2010; Čermák, 2012) sestaven seznam českých místních příslovčí, která by potenciálně mohla vytvářet spolu s primárními předložkami nějaké ustálené gramatické kombinace. V současné době tento pracovní seznam obsahuje přes stovku jednotek (včetně fonetických variant) a vypadá následovně: *blízko, blíž, blíže, nablízko, nablízku, poblíž, poblíže, daleko, nedaleko, opodál, dál, dále, zdaleka, vysoko, zvysoka, výše, výš, nízko, níže, níž, hluboko, zhluboka, hlouběji, vpravo, napravo, doprava, zprava, odprava, vlevo, nalevo, doleva, zleva, odleva, vpředu, vepředu, odpředu, dopředu, zpředu, zepředu, napřed, vpřed, vzadu, pozadu, dozadu, nazad, zezadu, odzadu, zadem, nahoře, nahoru, horem, odshora, shora, dole, dolů, dolem, odzdola, zezdola, zdola, uvnitř, dovnitř, zevnitř, vevnitř, uprostřed, vprostředku, doprostřed, zprostřed, vprostřed, středem, zprostředka, doprostředka, stranou, bokem, okrajem, zkraje, z boku, naproti, vstříc, napříč, navrch, naurchu, svrchu, vrchem, zespodu, zespoda, zespod, vespod, vespodu, odspodu, naspodu, naspod, dospodu, dospod, spodem, někde, všude, odevšad, odevšud, někam, někudy, kudysi, kudykoli, kudykoliv, kdesi, kdekoli, kdekoliv, kamsi, kamkoli, kamkoliv, odkudsi, odněkud*.

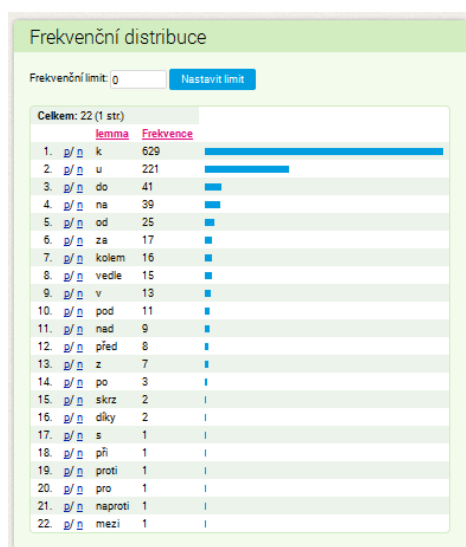
Dále byly vyhledány výskyty daných příslovčí ve spojení s předložkami v korpusu SYN2010.¹ Korpusový manažer Kontext nabízí k tomuto vyhledávání dvě možnosti – vyhledávání ve specifikovaném kontextu (1 token vpravo, předložka); v tomto případě bude konkordance obsahovat všechny výskyty tohoto příslovce s jakoukoliv předložkou v pravé pozici a soubor konkrétních kombinací s jejich frekvencí a dalšími statistickými hodnotami je možné nalézt v záložkách Frekvenční distribuce (Vlastní) a Kolokace; nebo vyhledávání v nespecifikovaném kontextu, tj. budou v konkordanci zobrazeny všechny výskyty hledaného příslovce v korpusu a dále, opět v záložkách Frekvenční distribuce (Vlastní) a Kolokace, je možné se podívat na to, jaké výrazy se nejčastěji vyskytují v pravém kontextu tohoto příslovce (jedna pozice vpravo) a jaké statistické hodnoty tyto kombinace mají. V daném

¹ *Český národní korpus – SYN2010* (2010). Praha: Ústav Českého národního korpusu FF UK v Praze. Dostupný z WWW: <<http://www.korpus.cz>>.

případě však budou do získaných přehledných tabulek zařazeny také kombinace příslovcí a interpunkčních znamének a jiných slov, což by náš výzkum zbytečně zatížilo. Většinou jsme využívali první z výše uvedených způsobů, který byl k danému výzkumu nejrelevantnější.

Další otázkou bylo nastavení frekvenčního limitu pro vyhledávání kombinací *Adv + Prep*. Na začátku jsme zkusili nedávat žádný limit, abychom se podívali na celkový souhrn vyskytujících se kombinací každého daného místního příslovce s jakýmikoliv předložkami. Pro každé příslovce jsme získali frekvenční seznamy takovýchto kombinací. „Partnerské“ předložky byly vyhledány jako lemmata, aby se v korpusových statistikách nerozlišovaly vokalizované a nevokalizované varianty stejné předložky. Jako příklad uvedeme frekvenční seznam kombinací předložek s příslovcem *blízko* (viz obrázek 1).

Obrázek 1: Frekvenční seznam kombinací předložek s příslovcem *blízko*.



V korpusu SYN2010 se vyskytuje 22 kombinací daného příslovce s předložkami. Svou frekvencí mezi ostatními výrazně vynikají dvě z těchto kombinací. Mají frekvenci více než 100 výskytů. Jsou to kombinace *blízko k* a *blízko u*.

Dále jsme se ve výzkumu zaměřili na zpracování právě nejfrekventovanějších kombinací místních příslovcí a předložek, které dle našich předpokladů jako víceslovné předložky fungují a jsou pro tuto skupinu předložkových jednotek charakteristické. Tímto způsobem jsme pro hledané kombinace stanovili frekvenční limit od 100 výskytů s tím, že méně frekventované jednotky budou zpracovány v dalším výzkumu. Jedná se například o jednotky *dovnitř do* (94 výskytů), *napříč přes* (55 výskytů), *uprostřed mezi + 7* (69 výskytů) aj. V dané etapě výzkumu jsme tedy využili kvantitativní data (v tomto případě údaje o frekvenci) pro omezení objektu výzkumu.

Pracovní seznam nalezených kombinací s frekvencí od 100 výskytů obsahuje v současné době 74 jednotek (viz tabulka 1). V tabulce 1 je uvedena absolutní frekvence každé kombinace, tj. celkový počet jejich výskytů v korpusu SYN2010, a také relativní frekvence, tj. procento podílu výskytů dané kombinace na celkovém počtu výskytů příslovce zařazeného do její struktury. Jednotky v tabulce 1 jsou seřazeny sestupně dle absolutní frekvence.

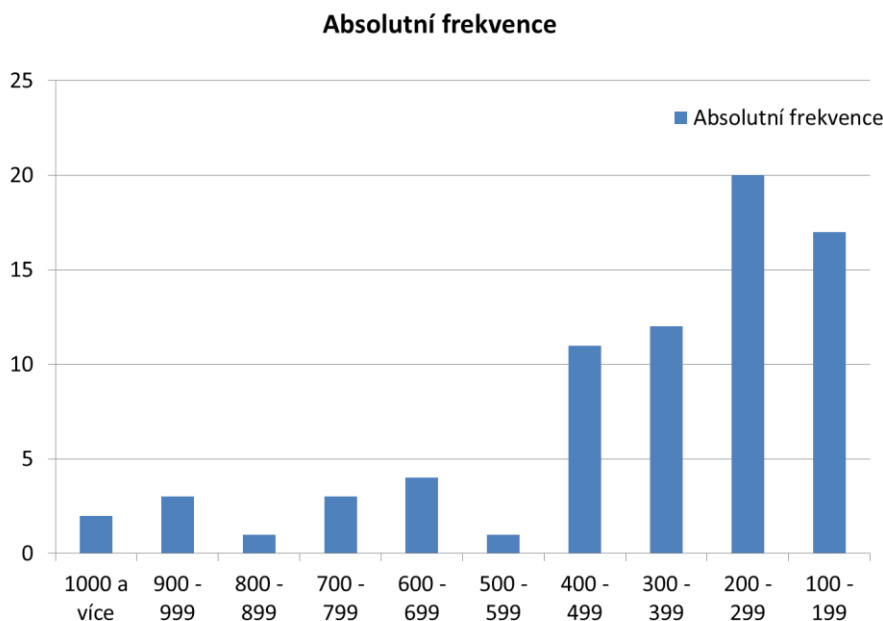
Tabulka 1: Kombinace místních příslovcí a předložek s frekvencí od 100 výskytů.

č.	Kombinace	Absolutní frekvence	Relativní frekvence (v %)
1	někde v + 6	2328	17,2
2	daleko od	1803	10,1
3	někde na + 6	944	6,9
4	dolů do	926	6,9
5	blíž k	900	25,5
6	dole v + 6	843	13,5
7	kdesi v + 6	783	34,8
8	hluboko do	779	22,4
9	dolů na + 4	700	5,2
10	vysoko nad + 7	684	16,8
11	blízko k	630	17,5
12	dolů po + 6	614	4,6
13	dolů k	600	4,5
14	někam do	589	10,9
15	dál do	486	1,4
16	dál v + 6	480	1,3
17	nahoře na + 6	476	8,1
18	všude kolem	473	4,2
19	blíže k	472	19,2
20	dále v + 6	460	2,1
21	hlouběji do	460	29,8
22	nahoře v + 6	442	7,5
23	dál na + 4	415	1,2
24	hluboko pod + 7	413	11,9
25	dále na + 4	403	1,8
26	daleko za + 7	380	2,1
27	všude v + 6	380	3,4
28	dole na + 6	377	6
29	daleko k	376	2,1
30	hluboko v + 6	369	10,6
31	vysoko v + 6	369	9,1
32	dál od	359	1
33	nedaleko od	356	26,3
34	vzadu v + 6	351	11,1
35	někam na + 4	349	6,5
36	odněkud z	336	34,4
37	všude na + 6	326	2,9
38	daleko do	290	1,6
39	někde u	286	2,1
40	dolů z	285	2,1
41	vzadu na + 6	280	8,9

42	kdesi na + 6	266	11,8
43	vysoko na + 6	265	6,5
44	dole pod + 7	261	4,2
45	nalevo od	259	25,7
46	dále do	258	1,2
47	dál po + 6	257	0,7
48	dole u	254	4,1
49	kamsi do	239	30,4
50	napravo od	231	22,2
51	někde mezi + 7	222	1,6
52	blízko u	221	6,1
53	daleko na + 6	218	1,2
54	nízko nad + 7	210	21,8
55	daleko v + 6	207	1,2
56	někde za + 7	205	1,5
57	vpravo od	203	7,2
58	vlevo od	197	6,9
59	stranou od	196	4,3
60	dál na + 6	192	0,5
61	daleko na + 4	183	1
62	všude po + 6	179	1,6
63	vzadu za + 7	178	5,6
64	daleko za + 4	176	1
65	vysoko nad + 4	162	4
66	dále na + 6	153	0,7
67	uvnitř v + 6	138	2,7
68	vysoko do	136	3,3
69	někam k	125	2,3
70	odkudsi z	119	31,2
71	hluboko pod + 4	115	3,3
72	dále od	109	0,5
73	daleko před + 7	104	0,6
74	dole na + 4	104	1,7

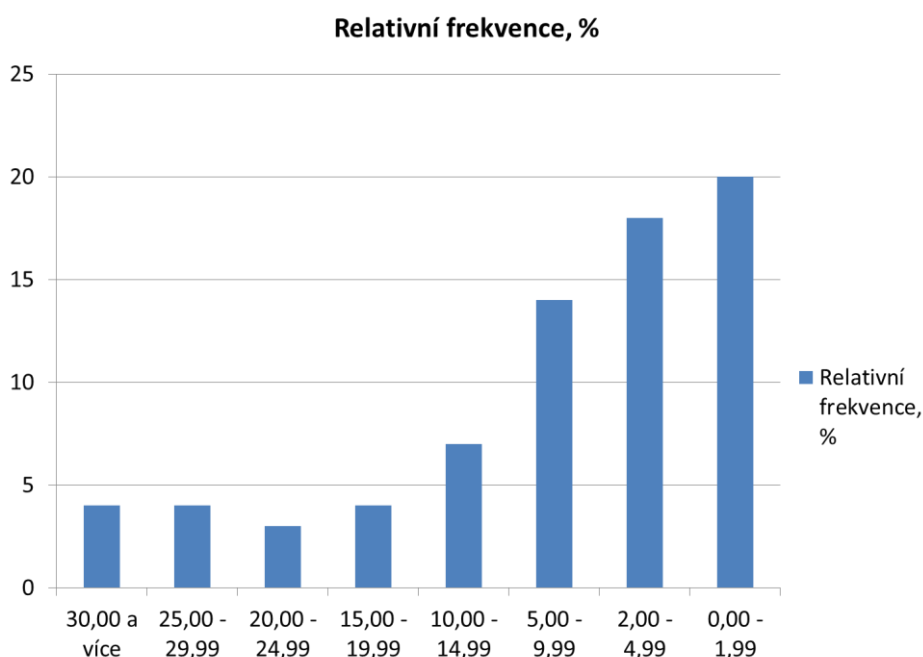
Tato data jsou přehledně zobrazena také v následujících grafech (viz grafy 1 a 2).

Graf 1: Absolutní frekvence kombinací místních příslovcí a předložek v korpusu SYN2010 s frekvenčním limitem od 100 výskytů.



Z grafu 1 vyplývá, že zkoumané jednotky mají většinou frekvenci 100–499 výskytů, 14 z nich má vyšší frekvenci. Na druhé pozici v tomto frekvenčním seznamu je jednotka *daleko od*, která se již bez komentáře objevovala v soupisu místních předložek uvedeném v PMCČ (Karlík et al., 1995). Srov. s jednotkou *stranou od*, která je v odborné literatuře evidována jako předložka (viz výše) a která má v daném seznamu 59. pozici.

Graf 2: Relativní frekvence kombinací místních příslovcí a předložek v korpusu SYN2010 s frekvenčním limitem od 100 výskytů (procento podílu výskytů kombinací na celkovém počtu výskytů příslovcí).



Graf 2 ukazuje, že zhruba pro desítku hledaných kombinací je podíl jejich výskytů na celkovém počtu výskytů příslušného příslovce značný – přes 20 %, což také může svědčit o jejich nenáhodnosti. Jsou to jednotky: *kdesi v + 6*, *odněkud z*, *odkudsi z*, *kamsi do*, *hlouběji do*, *nedaleko od*, *nalevo od*, *blíž k*, *hluboko do*, *napravo od*, *nízko nad +7*, srov. *daleko od*, které má relativní frekvenci 10,1 %, a *stranou od* s relativní frekvencí 4,3 %.

Domníváme se, že údaje o absolutní a relativní frekvenci zkoumaných jednotek tomuto výzkumu značně prospívají, jsou vhodné zejména pro objevování a výběr jádra zkoumaného jevu. Na tomto jádru je poté možné začít provádět kvalitativní analýzu.

Další přínosnou funkcí korpusu je vyhodnocení nalezených kombinací jako kolokací, a to pomocí asociačních měr. Na statistické míry zkoumaných kombinací *Adv + Prep* se můžeme podívat v záložce Kolokace. Vybereme si pro analýzu dvě z nejpoužívanějších měr – T-score, která poukazuje na pravidelnost a ustálenost kolokace (pro gramatické jevy je považována za nejrelevantnější), a MI-score, která vyjadřuje asociační sílu kolokace a je citlivá na spíše výjimečné, málo frekventované kolokace. Jakým způsobem jsou v korpusu zobrazeny statistické hodnoty hledaných kombinací *Adv + Prep*, ukážeme na příkladu příslovce *blízko* (viz obrázek 2). Předložkové kolokáty slova *blízko* jsou v této tabulce seřazeny sestupně podle hodnoty T-score.

Obrázek 2: Statistické míry kombinací předložek s příslovcem *blízko*.

Kolokace				
Celkem: 22 (1 str.)				
		Frekvence	T-score	MI
1. p/n	k	629	24.887	7.021
2. p/n	u	221	14.752	7.024
3. p/n	do	41	5.472	2.782
4. p/n	od	25	4.546	3.462
5. p/n	kolem	16	3.878	5.034
6. p/n	vedle	15	3.814	6.027
7. p/n	na	39	3.675	1.281
8. p/n	za	17	3.274	2.280
9. p/n	pod	11	3.112	4.017
10. p/n	nad	9	2.779	3.761
11. p/n	před	8	2.347	2.555
12. p/n	skrz	2	1.386	5.649
13. p/n	díky	2	1.243	3.047
14. p/n	naproti	1	0.954	4.438
15. p/n	proti	1	0.468	0.910
16. p/n	po	3	0.007	0.006
17. p/n	mezi	1	-0.016	-0.023
18. p/n	z	7	-0.222	-0.116
19. p/n	při	1	-0.339	-0.421
20. p/n	v	13	-2.087	-0.659
21. p/n	pro	1	-2.220	-1.687
22. p/n	s	1	-7.203	-3.036

V tomto korpusovém výpočtu statistických měr však nejsou kolokující předložky rozlišovány dle pádů, které řídí. To může ovlivňovat správnost získaných hodnot pro kombinace, ve kterých předložka řídí několik pádů (např. *pod*, *nad*, *před*, *v* aj.).

Vzhledem k tomu jsme pro další statistickou sondu udělali z již získaného souboru nejfrekventovanějších kombinací *Adv + Prep* (viz výše) a také kombinací méně frekventovaných (které jsou v daném příspěvku ponechány stranou) výběr z 25 jednotek obsahujících pouze jednovalenční předložky.

Získané statistiky těchto jednotek jsou zobrazeny v tabulce 2 (jednotky jsou seřazeny sestupně dle hodnoty T-score) a v tabulce 3 (jednotky jsou seřazeny sestupně dle hodnoty MI-score). Kurzivou jsou ke srovnání zvýrazněny jednotky, které již byly v odborné literatuře jako víceslovné předložky zmíněny.

Tabulka 2: Statistické míry vybraných kombinací místních příslovčí a předložek (jednotky seřazeny dle T-score).

č.	Kombinace	Frekvence	T-score	MI-score
1	<i>daleko od</i>	1803	42.259	7.710
2	blíž k	900	29.850	7.647
3	dolů do	927	29.734	5.416
4	hluboko do	779	27.462	5.961
5	blízko k	630	24.887	7.021
6	někam do	589	23.948	6.240
7	všude kolem	474	21.736	9.279
8	blíže k	472	21.602	7.463
9	hlouběji do	461	21.319	7.143
10	nedaleko od	356	18.817	8.529
11	odněkud z	336	18.185	6.977
12	někde u	286	16.449	5.194
13	<i>nalevo od</i>	259	16.051	8.559
14	dole u	245	15.446	6.245
15	kamsi do	239	15.300	6.602
16	<i>napravo od</i>	231	15.158	8.531
17	blízko u	221	14.752	7.024
18	vpravo od	203	14.170	7.511
19	vlevo od	197	13.953	7.413
20	<i>stranou od</i>	196	13.936	7.767
21	odkudsi z	119	10.815	6.856
22	vysoko do	136	10.809	3.773
23	naproti přes	108	10.375	9.266
24	dovnitř do	99	9.747	5.613
25	napříč přes	59	7.663	8.733

Tabulka 2 prokazuje, že hodnota T-score je přímo úměrná frekvenci kombinací.

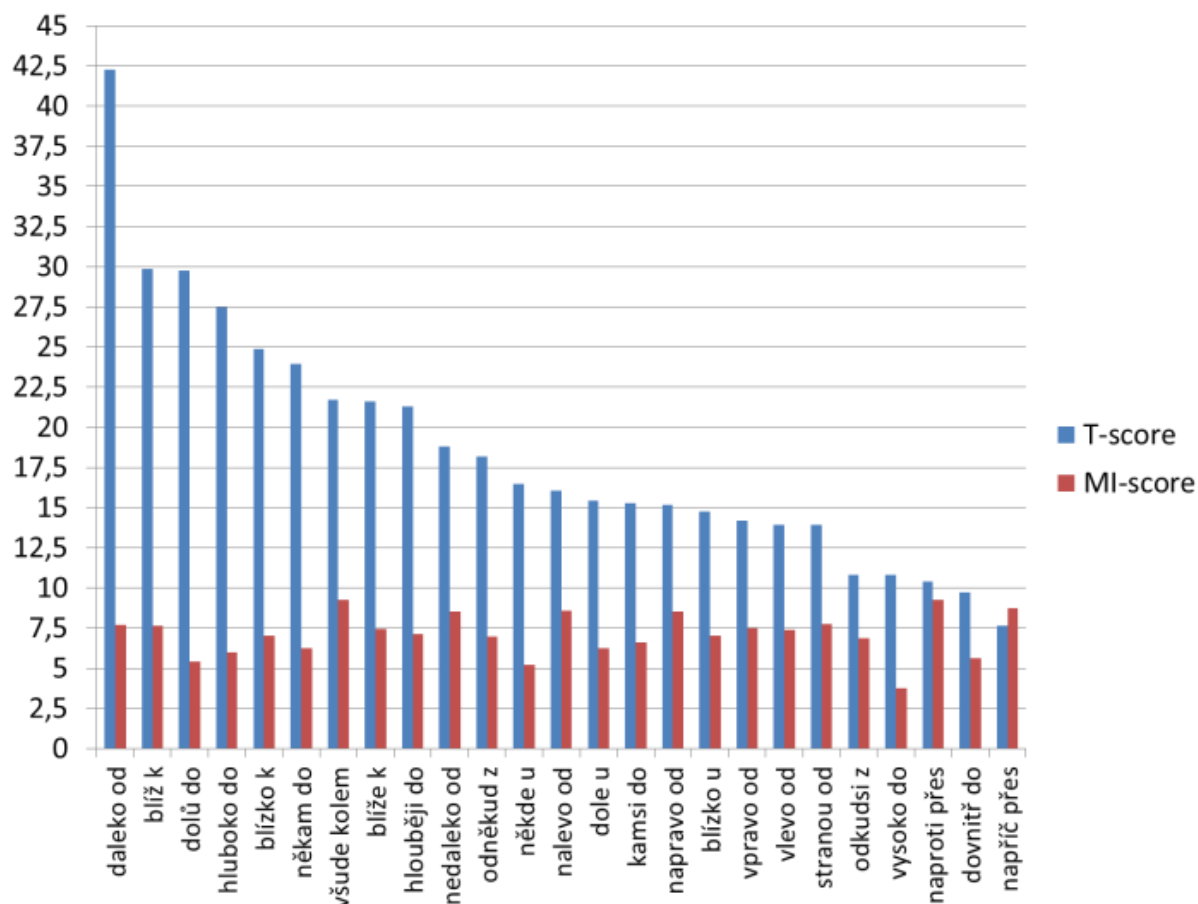
Tabulka 3: Statistické míry vybraných kombinací místních přísloví a předložek (jednotky seřazeny dle MI-score).

č.	Kombinace	Frekvence	T-score	MI-score
1	všude kolem	474	21.736	9.279
2	naproti přes	108	10.375	9.266
3	napříč přes	59	7.663	8.733
4	<i>nalevo od</i>	259	16.051	8.559
5	<i>napravo od</i>	231	15.158	8.531
6	nedaleko od	356	18.817	8.529
7	<i>stranou od</i>	196	13.936	7.767
8	<i>daleko od</i>	1803	42.259	7.710
9	blíž k	900	29.850	7.647
10	vpravo od	203	14.170	7.511
11	blíže k	472	21.602	7.463
12	vlevo od	197	13.953	7.413
13	hlouběji do	461	21.319	7.143
14	blízko u	221	14.752	7.024
15	blízko k	630	24.887	7.021
16	odněkud z	336	18.185	6.977
17	odkudsi z	119	10.815	6.856
18	kamsi do	239	15.300	6.602
19	dole u	245	15.446	6.245
20	někam do	589	23.948	6.240
21	hluboko do	779	27.462	5.961
22	dovnitř do	99	9.747	5.613
23	dolů do	927	29.734	5.416
24	někde u	286	16.449	5.194
25	vysoko do	136	10.809	3.773

Z tabulky 3 vyplývá, že většina zkoumaných kombinací má vysokou hodnotu MI-score (za relevantní pro systémovou kolokaci je považována hranice $MI = 7$ (Kocěk et al., 2000)), což tyto kombinace opravňuje považovat za frazeologické kolokace. Srov. začlenění víceslovných předložek do frazémů u Renaty Blatné (2004).

Relace mezi získanými hodnotami T-score a MI-score jsme znázornili v grafu 3.

Graf 3: Míry T-score a MI-score vybraných kombinací místních příslovčí a předložek.



Z grafu 3 vyplývá, že asociační síla zkoumaných kombinací nezávisí na pravidelnosti a frekvenci jejich používání a většinou je dosti značná.

Získané statistické hodnoty ukazují na pravidelnost a nenáhodnost daných souvýskytů, což umožňuje předpoklad, že se dané víceslovné útvary mohou vyskytovat ve funkci stejné s funkcí jednoslovné předložky. Pro ověření tohoto předpokladu je však potřeba provést další kvalitativní analýzu nalezeného souboru jednotek.

Závěrem lze shrnout, že v daném výzkumu byly statistické funkce korpusu využity především ke stanovení přibližného souboru jednotek, které předtím neměly důkladný lingvistický popis. Relevantnost získaných kvantitativních dat však bude možné nejobektivněji posoudit během následující kvalitativní analýzy.

Literatura:

- BLATNÁ, Renata (2004): Využití statistických metod při popisu neverbálních kolokací. *Slovo a slovesnost*, 65, s. 24–52.
- BLATNÁ, Renata (2006): *Víceslovné předložky v současné češtině*. Praha: Nakladatelství Lidové noviny.
- CVRČEK, Václav et al. (2010): *Mluvnice současné češtiny 1: Jak se píše a jak se mluví*. Praha: Karolinum.
- ČERMÁK, František – HRONEK, Jiří – MACHAČ, Jaroslav (eds.) (1988): *Slovník české frazeologie a idiomatiky: Výrazy neslovesné*. Praha: Academia.

- ČERMÁK, František (2012): *Morfematika a slovtvorba češtiny*. Praha: Nakladatelství Lidové noviny.
- KANJUŠKEVIČ, Maryja (2008): *Belaruskija prynazoŭniki i ich analahi: Hramatyka reál'naha ŭžyvannja: Materyjaly da sloŭnika*. Hrodna: HrDU.
- KOCEK, Jan – KOPŘIVOVÁ, Marie – KUČERA, Karel (eds.) (2000): *Český národní korpus: Úvod a příručka uživatele*. Praha: FF UK v Praze.
- KROUPOVÁ, Libuše (1985): *Sekundární předložky v současné spisovné češtině*. Praha: Ústav pro jazyk český ČSAV.
- MČ2: KOMÁREK, Miroslav – KOŘENSKÝ, Jan – PETR, Jan – VESELKOVÁ, Jarmila (eds.) (1986): *Mluvnice češtiny 2: Tvarosloví*. Praha: Academia.
- PMČ: KARLÍK, Petr – NEKULA, Marek – RUSÍNOVÁ, Zdenka (eds.) (1995): *Příruční mluvnice češtiny*. Praha: Nakladatelství Lidové noviny.
- ŠUBA, Pavel (1993): *Tlumačal'ny sloŭnik belaruskich prynazoŭnikaŭ*. Minsk: Narodnaja asveta.
- VSEVOLODOVA, M. – VINOGRADOVA, J. – KLOBUKOV, J. – KUKUŠKINA, O. – POLIKARPOV, A. – ČEKALINA, V. (v tisku): *Materialy k slovarju „Predlogi i sredstva predložnogo tipa v ruskom jazyke: Funkcional'no-kommunikativnyje aspekty real'nogo upotreblenija“*. Moskva.