Univerzita Palackého v Olomouci

Filozofická fakulta

Katedra Bohemistiky

Česká filologie

Základní statistické metody pro vytěžování jazykových dat a jejich možnosti pro analýzu textů

Bakalářská práce

Olomouc 2018

Vedoucí bakalářské práce:

PhDr. Petr Pořízka, Ph.D.

Vypracovala:

Michela Láníčková

Prohlašuji, že jsem závěrečnou práci vypracovala samostatně a v seznamu literatury jsem uvedla veškeré informační zdroje, které jsem použila.

V Olomouci dne 3. 5. 2018 ……

Na tomto místě bych ráda poděkovala vedoucímu své bakalářské práce PhDr. Petru Pořízkovi, Ph.D. za odborné konzultace.

Obsah

1 Úvod

Téma bakalářské práce jsem si zvolila Základní statistické metody pro vytěžování jazykových dat a jejich možnosti pro analýzu textů. V této práci budu pracovat s Českým národním korpusem a vycházet z oboru korpusová lingvistika.

Korpusová lingvistika je podle Petra Pořízky[1] jednou z nejdůležitějších metodologií pro výzkum současného jazyka. Korpus definuje [2] jako soubor elektronicky zpracovaných textů, které mohou být anotovány a slouží především ke zkoumání jazyka nebo literárnímu výzkumu. Velkou předností korpusu je schopnost podávat frekvenci jevů a jejich úzu.

V současnosti existuje řada korpusů,[3] které nám svou velikostí nebo autentičností textu (přepsané zvukové záznamy) umožňují efektivní zkoumání jazyka. K práci s korpusy nám slouží nejčastěji korpusové manažery neboli konkordanční nástroje, to jsou speciální aplikace, jež umožňují vyhledávání v korpusových datech. Korpusové manažery nám umožňují vyhodnocování základních frekvenčních a distribučních statistik, případně složitějších statistických analýz, vyhledávání konkrétních konkordancí, kolokací. Příkaz k vyhledávání zadáváme ve speciálním dotazovacím jazyce, nejčastěji v CQL (Corpus Query Languge). Pokud chce uživatel pracovat s korpusem na základní úrovni, je pro něj alespoň částečná znalost dotazovacího jazyka nutností.

Korpusová lingvistika se ve světě začala vyvíjet od šedesátých let 20. století, kdy lingvisté H. Kučera a W. N. Francis z Brownovy univerzity vytvořili korpus soudobé pouze psané americké angličtiny.[4] U nás se korpusová lingvistika začala rozvíjet roku 1994, kdy byl založen Ústav Českého národního korpusu na Filozofické fakultě Univerzity Karlovy v Praze (UČNK).

V teoretické části se zaměřím na vysvětlení základních pojmů, s kterými budu dále pracovat, především pojmy: Český národní korpus, kolokace a jejich členění

[1] Pořízka P., Tvorba korpusů a vytěžování jazykových dat: metody, modely nástroje, Olomouc 2014, str. 9.
[2] Tamtéž, str. 10.
[3] Např.: Korpus DIALOG, Česká elektronická knihovna, Korpus českého verše, Vokabulář webový, Pražský závislostní korpus aj.
[4] Computation Analysis of Present-Day American English.

a asociační míry. V praktické části provedu analýzu základních asociačních měr na šesti vybraných lemmatech.[5]

2 Teoretická část

2.1 Český národní korpus

Český národní korpus (ČNK) je projekt zaměřený zejména na budování rozsáhlého počítačového korpusu především psané češtiny. Velikost ČNK není stálá, objem dat v korpusu stále narůstá, celkový rozsah textů přesahuje 3,6 miliardy slov v českých jednojazyčných a 1,5 miliardy v cizojazyčných paralelních korpusech.[6]

Počítačová databáze Českého národního korpusu je typologicky rozřazena do jednotlivých korpusů. Rozděluje korpusy psané a mluvené. Synchronní korpusy jsou budovány jako reprezentativní ukázka jazyka za určité období, naproti tomu diachronní korpusy se snaží o zachycení jazyka v různých vývojových fázích, korpus DIAKORP obsahuje texty, jež svým rozsahem pokrývají sedm století. Nereferenční korpusy jsou takové, které se pravidelně aktualizují, a stejný dotaz zadaný v různém období tak může vygenerovat jiné výsledky, referenční korpusy jsou naopak celistvé, stejný příkaz pak generuje stejné výsledky nezávisle na období, kdy byl příkaz zadán. Cizojazyčné korpusy mohou zahrnovat texty rozličných jazyků, např. korpus InterCorp obsahuje více než třicet jazyků, nebo mohou být tvořeny jen z jednoho nečeského jazyka, např. korpus deWaC je webový korpus němčiny.

2.2 Kolokace

Pojem kolokace stále ještě nemá jednotnou definici, a to i přes to, že její studium proniká stále více do centra zájmu v oblasti lingvistiky. V předkládané práci budeme vycházet především z vymezení stanoveného F. Čermákem, viz níže, ale pro úplnost jsme se rozhodli alespoň nastínit i některá jiné vymezení tohoto pojmu.

[5] Souhrn všech tvarů lexikální jednotky.
[6] Cvrček – Richterová: Příručka ČNK, 2013, online.

Kolokaci jako pojem do jazykovědy zavedl a rozvedl J. R. Firth,[7] na něj pak dále navázal např. J. Sinclair. Na české území tento pojem poprvé uvedl F. Čermák.[8] Kolokací označuje kombinaci dvou i více slovních textových tvarů, která je závislá na sémantických pravidlech, ta řídí kombinaci slov a kolokace ukazují, co je běžné a všední, nebo co už je naopak netypické.

Tato podmínka, totiž že slova na sobě musejí být sémanticky závislá, tedy vylučuje, aby bylo možno považovat za kolokaci jakákoliv sousedící slova – taková slova označujeme jako n-gram, přičemž pod tento pojem spadají bigramy, trigramy atd. Tyto dva pojmy (kolokace a n-gramy) bychom tedy neměli zaměňovat i přesto, že jsou si svým významem blízké a kolokace může být považována za součást n-gramu.

Sousedství slov se sémantickou vazbou se může vyskytovat až do vzdálenosti pěti slov od sebe, ale většina spojení se nachází spíše v těsné blízkosti. Mezi těmito slovy ale musí fungovat pravidlo sémantické slučitelnosti, tedy kompatibility. Čermák uvádí jako příklad slova *zuřivě spí*. Taková slova nejsou kompatibilní, protože nemají význam, ale naopak *klidně spí* už za kompatibilní označit můžeme.

Pokud u slova *spí* vyčerpáme celý souhrn jeho kolokací, mluvíme o kolokačním paradigmatu, to je pro toto slovo jedinečné a nebude se shodovat s jiným slovem. Slovo, které se připojuje k ústřednímu slovu, v našem případě tedy adjektivum *klidně*, označujeme jako kolokát.

Sinclair,[9] ze kterého Čermák částečně vychází, za kolokaci považuje slova, jež jsou nedaleko od sebe.[10] Český národní korpus[11] vymezuje kolokaci jako podmíněnost sémantické kompatibility členů, kde význam členů dohromady často přesahuje jejich význam zvlášť. Zde jako příklad uvádí spojení *cestovní ruch*, v němž se význam členů stojících izolovaně liší od významu, pokud stojí slova v těsné blízkosti, přičemž použití jednoho členu předznamenává použití členu druhého. V této

[7] Frith J. R., 1957, Models of meaning. In Papers in Linguistics 1934-51, Oxford U. P., London.
[8] Čermák F., J. Holub, 2005 (1982), Syntagmatika a paradigmatika českého slova. 1. Valence a kolokabilita. Karolinum, Praha.
[9] Sinclair J., Corpus, Concordance, Collocation., 1991, str. 70.
[10] Hranici zvolil maximálně pět slov napravo a pět slov nalevo od klíčového lemmatu.
[11] Cvrček – Richterová: Příručka ČNK, 2013, online.

souvislosti mluví Václav Cvrček o principu předurčeného výběru.[12]

V souvislosti s kolokací je důležité zmínit i koligaci, jež je někdy považována[13] za specifický druh kolokace. O koligaci mluvíme, pokud je namísto lexikálně-sémantických vztahů zohledňován vztah mezi lexikální jednotkou a gramatickou kategorií.

2.3 Členění kolokace

Členění kolokace je podobně nejednotné, jako snahy o její definování. Např. F. Čermák[14] člení lexikální kombinace v textu na (A) systémové, (B) textové a (C) textové-systémové

(A) Systémové

1 pravidelné a: **termínové kolokace** (cestovní kancelář)

b: **propriální kolokace** (Kanárské ostrovy)

2 nepravidelné **idiomatické kolokace** (ležet ladem)

(B) Textové

3 pravidelné a: **běžné kolokace** (letní dovolená)

b: **analytické kombinace tvarů** (byl zapsán)

4 nepravidelné a: **individuální metaforické kolokace** (virové hrátky)

b: **náhodné kombinace sousední** (*vývody* vzduchotechniky uvnitř)

c: **jiné kombinace** (blábol)

(C) Textové-systémové

5 běžné kolokace uzuální (prát prádlo)

Klasifikace F. Čermáka je jen jednou z mnoha a důležité je upozornit, že toto vymezení slouží spíše jako návod pro uživatele, jak na kolokace nahlížet, neboť hranice mezi jednotlivými typy nejsou ostré, zvláště pak mezi typy A1a a B3a. O něco jednodušší je třídění kolokátů na (A) volné kombinace, (B) těsné kombinace a (C) vlastní kombinace:[15]

[12] Tamtéž.
[13] Tamtéž.
[14] Čermák F., Syntagmatika slovníku: typy lexikálních kombinací. In Čeština – univerzália a specifika 3, eds. Z. Hladká, P. Karlík, 2001, str. .223-232.
[15] Tamtéž str. 28.

(A) Volné kombinace (mýt auto, mýt nádobí)

(B) Těsné kombinace (parkovat auto, pasterizovat mléko)

(C) Vlastní kombinace (Kanárské ostrovy, kyselina sírová)

Za volné kombinace se považují takové, kde kolokáty jsou snadno nahraditelné, např. verbum *mýt* se pojí s celou škálou jiných slov. Těsné kombinace mají funkční pole o něco omezenější, sémanticky nedává smysl například *pasterizovat nádobí*, protože pasterizace je proces, který se používá při konzervaci potravin. Sloveso se tedy bude převážně pojit s nějakou potravinou (mléko, pivo atd.). Vlastní kombinace jsou pak takové, které mají svůj vlastní specifický a omezený výběr kolokátů.

Dále bychom podle Sinclaira [16] měli rozlišovat kolokáty vzestupné a sestupné. Sestupným kolokátem označuje slova, kde druhé vyhledané slovo má oproti prvnímu nižší frekvenci, naopak vzestupný kolokát je ten, u kterého první slovo má oproti následujícímu nižší frekvenci. Ze statistického hlediska se Sinclair domnívá, že vzestupné kolokáty jsou méně významné a spíše upřesňují gramatický rámec hledaného slova, zatímco vzestupné kolokáty přispívají k sémantické analýze slova.

2.4 Asociační míry

Kolokace se vyhledávají za pomoci statistických měr, a ty přiřazují kolokacím číselnou hodnotu. Asociační míry podle ČNK [17] pracují s frekvencí celé kolokace, jednotlivých členů a velikosti korpusu. Výsledná hodnota n-gramů pak vyjadřuje míru asociace, přičemž tato hodnota může být i záporná, pokud je mezi členy vztah tzv. negativní asociace, tzn. členy se vzájemně odpuzují. Hodnota asociační míry jednoho n-gramu není srovnatelná s číselnou hodnotou jiného n-gramu, k takovémuto srovnání slouží převedení číselných hodnot na pořadí tedy ranků, [18] které je uspořádané podle číselných hodnot dané asociační míry.

Tyto hodnoty by měly být zpravidla získávány z korpusu, který má širokou základnu, co se týče jeho velikosti, aby výsledky mohly být objektivní a neměnné. Studium

[16] Sinclair J., Corpus, Concordance, Collocation. Oxford U.P., 1991 , str. 115-116.
[17] Cvrček – Richterová: Příručka ČNK, 2013, online.
[18] Rank představuje relativizaci frekvence, kdy rank 1 má nejvyšší frekvenci a rank n, kde N je celkový počet položek v seznamu, jevu s frekvencí nejnižší.

kolokací vyžaduje využití počítače. Toto tvrzení ostatně předkládá i Jeremy Clear[19] ve své studii, kde píše, že množství lexikálních jednotek je tak velké, že je až nemožné, aby takovéto studium prováděl jedinec. Zároveň u takovéhoto zkoumání nastává větší pravděpodobnost chybného zpracování. Je také nutné nespoléhat se jen na počítačové zpracování dat a výsledky ručně protřídit.

M. Křen [20] upozorňuje, že většina takovýchto měr byla testována na anglickém jazyce. Angličtina má v rámci tradiční Skaličkovy jazykové typologie nejblíže k tzv. analytickému pólu,[21] pro který je mimo jiné typická vysoká polysémie slov. Čeština má naopak blíže k flektivním jazykům (resp. k flektivnímu jazykovému typu), a je pro ni tedy typická diferenciace morfému a slova. Hodnoty, které udávají statistické míry pro anglický jazyk, se mohou odlišovat od těch, které jsou výsledkem zkoumání na česky psaném korpusu. Z tohoto důvodu je nezbytné, abychom jednotlivé míry otestovali a popsali i na českém jazyce.

Statistických testů pro vyhledávání n-gramů je velké množství. Nejčastěji se však používají testy t-score, MI-score, MI3-score, Dice, logDice, logLikelihood, Chi-squared. Z nich nejtypičtější jsou první dva jmenované, tedy t-score a MI-score.
P. Pořízka [22] největší rozšířenost těchto dvou měr dokazuje na výčtu měr u jednotlivých konkordančních nástrojů.
Manatee/Bonito (v1) – MI-score, t-score
AntConc – MI-score, t-score, logLikehood, Chi-squared
Xaira – MI-score, z-score
Manatee/(No|Word) Sketch Engine (v2)/ KonText – MI-score, t-score, MI3-score, logLikelihood, minimální citlivost, logDice, MI.log_f, relative freq
Všechny tyto nástroje tedy využívají míry MI-score a t-score, až na nástroj Xaira, jenž využívá z-score, který si je svou povahou podobný s t-score mírou.

[19] Z Firthovských principů: komputační nástroje pro studium kolokace (překlad stati J. Cleara). In: Studie z korpusové lingvistiky, Praha, Karolinum 2000, str. 495-513.
[20] Křen M., Kolokační míry a čeština: srovnání na datech Českého národního korpusu, In: Kolokace Čermák F. a kol., 2006, str. 224.
[21] Skalička, Vladimír. Typ češtiny, 1951, str. 10.
[22] Pořízka P., Tvorba korpusů a vytěžování jazykových dat: metody, modely nástroje, Olomouc 2014, str. 40.

2.4.1 MI-score

Zkratka MI zde označuje mutual information tedy výsledek vzájemné informace.

$$I(xy) = \log_2 \frac{p(xy)}{p(x)p(y)}$$[23]

OBR. Č. 1: VZOREC PRO VÝPOČET MÍRY MI-SCORE 1

Kde p označuje pravděpodobnost jevu, takže p(x) chápeme, jako pravděpodobnost jevu x, p(y) je pravděpodobnost jevu y a p(xy) je pravděpodobnost, že jevy x a y nastanou současně.

Pravděpodobnost dále pak počítáme jako frekvenci slov vydělenou velikostí korpusu, tady máme na mysli počtem tokenů [24] v korpusu. Tedy po dosazení do činitele dostáváme $p = f(xy)/N$ a to stejné provedeme i ve jmenovateli $p = f(x)/N$ a $p = f(y)/N$.

Z toho vyplývá vzorec:

$$MI(xy) = \log_2 \frac{\frac{f(xy)}{N}}{\frac{f(x)}{N}\frac{f(y)}{N}} = \log_2 \frac{N f(xy)}{f(x) f(y)}$$[25]

OBR. Č. 2: VZOREC PRO VÝPOČET MÍRY MI-SCORE 1

Test tedy měří podíl pravděpodobnosti výskytu slov blízko sebe a výskytu každého členu nezávisle na sobě. Platí zde, že nejvyšších hodnot dosahují slova s nižší frekvencí. Míra MI-score je vhodná pro hledání neobvyklých kolokací a také slouží k vyhledávání jádra frazémů v rámci celé škály kolokace.

Nevýhodou této míry je, že její výsledky jsou značně ovlivněny frekvencí jednotlivých slov, a je tedy u této míry potřeba nastavit spodní hranici frekvence hledaného slova. Jinak se může stát, že mezi výslednými kolokacemi se objeví členy, kde jejich vzájemný výskyt je jen chybné použití slov, nebo výsledek užívání idiolektu.

[23] Cvrček – Richterová: Příručka ČNK, 2013, online.

[24] Tokenem označujeme nejmenší jednotku textu, která je oddělená mezerou.

[25] Cvrček – Richterová: Příručka ČNK, 2013, online.

2.4.2 T-score

T-score je statistická metoda, označována jako míra kontrastu, která se využívá na testování hypotéz. Testuje, zda zjištěné počty výskytu jednotlivých slov a jejich členů odpovídají náhodnému rozložení slov v korpusu. Vzorec pro tuto míru vypadá následovně:

$$T = \frac{\left(f(x,y) - \frac{f(x) \cdot f(y)}{N}\right)}{\sqrt{f(x,y)}}$$

[26]

OBR. Č. 3: VZOREC PRO VÝPOČET MÍRY T-SCORE

T-score pracuje se stejnými hodnotami jako MI-score, takže *f(x,y)* označuje frekvenci jevů *x* a *y*, *f(x)* frekvenci jevu *x* a *f(y)* frekvenci jevu *y*, *N* je počet tokenů korpusu, tedy celková velikost korpusu.

V čitateli se jedná o porovnání skutečné (naměřené) hodnoty, od které se odečítá hodnota očekávaná jevů x a y. Čitatel se pak podělí odmocninou skutečné hodnoty.
Petr Pořízka[27] vzorec interpretuje jako rozdíl mezi pozorovaným a předpokládaným. Míra zjišťuje, zda počet výskytů jednotlivých slov a jejich dvojic odpovídají náhodnému rozložení slov. Čím vyšší je hodnota t-score, tím nižší je pravděpodobnost, že se jedná o náhodný spolu výskyt členů. Tedy pokud členy mají naměřenou vysokou hodnotu t-score, je pravděpodobné, že se jedná o pevnější, ustálenější kombinaci slov.

Tento test je vhodný pro vyhledávání frekventovaných spojení, která jsou sémanticky pravidelná a svou povahou gramatická, vhodná pro vyhledávání koligací. Neměří sílu asociace hledaných výrazů, ale vyjadřuje míru přesvědčení, že mezi výrazy je nějaká pevná kolokační nebo koligační asociace. t-score oproti MI-score bere v úvahu četnost výskytu a nenadhodnocuje spojení, u nichž má člen nízkou frekvenci.

[26] Tamtéž.
[27] Pořízka P., Tvorba korpusů a vytěžování jazykových dat: metody, modely nástroje, 2014, str. 42.

2.4.3 Log likelihood

Další asociační míra, kterou se pokusíme v této práci přiblížit, je Log likelihood. Výsledky toho testu jsou sice ovlivněny počtem tokenů v korpusu (stejně jako u míry t-score), ale u míry Log likelihood velikost korpusu nehraje tak značnou roli při udávání výsledků, jako je tomu u míry t-score.[28] Míra reflektuje rozdíl mezi pozorovanými a očekávanými hodnoty.

$$LL(xy) = f(xy)\log(f(xy))+(f(x)-f(xy))\log(f(x)-f(xy))+(f(y)-f(xy))\log(f(y)-f(xy))+N\log N$$
$$+(N+f(xy)-f(x)-f(y))\log(N+f(xy)-f(x)-f(y))-f(x)\log(f(x))-f(y)\log(f(y))-(N-f(x))\log(N-f(x))-(N-f(y))\log(N-f(y))$$

OBR. Č. 4: VZOREC PRO VÝPOČET MÍRY LOG LIKELIHOOD 1[29]

Log Likelihood porovnává skutečný výskyt n-gramu s jeho předpokládaným výskytem. Tedy míra porovnává frekvenci sledovaných výpočtů, které očekáváme, že nastanou, pokud počet tokenů v n-gramu koresponduje s předpokládaným modelem. Za takovýto předpokládaný model je označovaná pravděpodobnost, že se dvě slova objeví vedle sebe.

M. Křen[30] ve své studii shrnul log Likelihood jako konzervativní vylepšenou verzi míry t-score, jehož extrémy má tendenci potlačovat a udávat významnější bigramy, míra tak silně nepreferuje bigramy s vysokou frekvencí, jak je tomu právě u míry t-score.

2.4.4 logDice

Další mírou, kterou budeme zkoumat, je míra logDice. Rozdíl oproti mírám, jež jsme zmiňovali výše, tedy MI-score, t-score a log likelihood je ten, že logDice nepracuje s velikostí korpusu. Míra je závislá jen na frekvenci slov *x* a *y* a na frekvenci bigramu *xy*. I zde uvádíme vzorec těchto dvou měr:

[28] Dunning T., Accurate Methods for the Statistics of Surprise and Coincidence, online.
[29] Cvrček – Richterová: Příručka ČNK, 2013, online.
[30] Křen M., Kolokační míry a čeština: srovnání na datech Českého národního korpusu, In: Kolokace Čermák F. a kol., 2006, str. 245.

[31]

$$logDice(xy) = 14 + \log_2 \frac{2f(xy)}{f(x) + f(y)}$$

OBR. Č. 5: VZOREC PRO VÝPOČET MÍRY LOGDICE

Původní míra Dice udává poměr bigramu *xy* a frekvenci slova *x* přičtenou k frekvenci slova *y*.[32] Jelikož frekvence bigramu *xy* bude vždy stejná jako frekvence slov *xy* nebo nižší, udává míra Dice nízké výsledky, její rozsah je v rozmezí 0, – 1, byla tedy upravena na míru logDice. Tato míra udává větší výsledné hodnoty, jejich rozsah je od mínus 14 do nekonečna.

Záporné hodnoty nastávají, pokud bigram není brán jako statisticky významná kolokace. Hodnota nula znamená, že byl nalezen méně než jeden výskyt bigramu *xy* pro 16 000 *x* nebo 16 000 *y*. Míry Dice a tedy i logDice, nejsou závislé na velikosti korpusu.

2.4.5 Minimální citlivost

Tato míra stojí na jednoduchém základu a udává poměrně zajímavé kolokace. I zde uvádíme její vzorec pro lepší představu toho, jak míra pracuje.

$$MS(xy) = \min\left(\frac{f(xy)}{f(x)}, \frac{f(xy)}{f(y)}\right)$$ [33]

OBR. Č. 6: VZOREC PRO VÝPOČET MINIMÁLNÍ CITLIVOST 1

Test je na výpočet daleko jednodušší oproti předchozím mírám, pracuje jen s frekvencí bigramu *xy,* frekvencí členu *x* a frekvencí členu *y*, míra tedy nebere ohled na velikost korpusu a pro každý člen se citlivost vypočítává zvlášť. Vztah minimální citlivosti je podmíněn pravděpodobností, jak často se členy vyskytnou v daném bigramu a jak často se členy vyskytnou celkově. Výsledné hodnoty se pohybují od nuly do jedné, hodnota jedna se vyskytuje, pokud člen *x* a člen *y* se vyskytnou

OBR. Č. 5: VZOREC PRO VÝPOČET MÍRY LOGDI 2

[31] Cvrček – Richterová: Příručka ČNK, 2013, online.
[32] Rychlý, P. A Lexicographer-Friendly Association Score.
[33] [33] Cvrček – Richterová: Příručka ČNK, 2013, online.

pouze a jenom současně, a to v bigramu *xy*, pokud tento výskyt nenastane ani jednou, hodnota míry se rovná nule.

T. Pedersen[34] v závěru své práce minimální citlivost popisuje jako míru se schopností dobře rozpoznat bigramy, které pojí jejich významová složka.

3 Praktická část

Pro zkoumání měr jsme se rozhodli vytvořit vlastní subkorpus, který jsme dostali z řady SYN2010. Náš subkorpus jsme vytvořili tak, aby byla ve vyváženém poměru odborná literatura, beletrie a publicistika. Rozložení korpusu SYN2010 není totiž ve vyváženém poměru a 40 % textů je beletristických na úkor odborné literatury. Velikost našeho vzorového subkorpusu je 1002397 slov,[35] takovou velikost pokládáme za přijatelnou, abychom vzniklé výsledky mohli považovat za dostatečně validní.

Subkorpus v ČNK lze vytvořit pomocí příkazu,[36] tuto možnost je dobré použít, pokud uživatel subkorpus chce využít jen jednorázově, nebo si může uživatel vytvořit trvalý subkorpus prostřednictvím položky v menu Korpusy → Vytvořit nový subkorpus. My jsme zvolili druhou možnost, která je pro naše účely vhodnější, protože subkorpus budeme potřebovat více než jednou.

Abychom mohli popsat povahu měr, co nejlépe, zvolili jsme si pro jejich zkoumání šest lemmat tří různých slovních druhů: substantivum, adjektivum a adverbium. Pro každý slovní druh jsme vybrali dvě lemmata, přičemž jedno reprezentuje lemmata s nízkou frekvencí a druhé s frekvencí vysokou. Jako zástupce substantiv jsme zvolili lemma *automobil* s frekvencí devadesát devět a lemma *pán* s frekvencí pět set sedmdesát tři. Za zástupce adjektiv jsme vybrali lemma *šťastný* s frekvencí devadesát šest a lemma *politický* s frekvencí dvě stě padesát čtyři. Jako zástupce adverbií jsme vybrali lemma *místy* s frekvencí devadesát šest a lemma *rychle* s frekvencí dvě stě dvacet sedm. Výsledné řady, co nám míra vygeneruje popíšeme. Pro přehlednost lemmata, která budou schopná tvořit kolokaci s naším KWIC

[34] Pedersen T., Dependent Bigram Identification, online.
[35] Viz přílohu č.1.
[36] Viz přílohu č. 2.

označíme jako kolokáty a vždy je uvedeme v kontextu. Pro zbylá lemmata, necháme označení lemma nebo jiné.

3.1 Lemma automobil

První námi zkoumané lemma je lemma *automobil*, které má v námi vytvořeném subkorpusu počet výskytů devadesát devět. Příkazem [37] jsme si jej vyhledali a následně vytvořili kolokaci. Kolokace → Vlastní, tento postup je pro všechna lemmata stejný, počet pozic nalevo od KWICu[38] jsme zadali mínus tři a napravo tři.

3.1.1 MI-score

Míra MI-score vygenerovala řadu o celkových šedesáti dvou pozicích, do první poloviny řady vygenerovala míra slušný výsledek, kde převažují autosémantika s frekvencí nižší nebo rovno čtyřem.

Výjimku tvoří kolokát[39] *nákladní* na druhé pozici s frekvencí osm s hodnotou 18.837, ve významu např.: *...na nákladních automobilech je vozí odběratelům a vysypou například do sklepa...* [40] Kolokát *užitkový* na třetí pozici s hodnotou 17.907 má frekvenci pět v kontextu ve větě např.: *...velké a menší firmy stále ještě považují osobní i užitkový automobil za jakýsi nadstandard a vytvářejí si podle Jaroslava Laura...*[41] a kolokát *osobní* s frekvencí třináct a hodnotou 16.653 na pozici osmé, například ve významu *...řidič osobního automobilu srazil tři dívky přebíhající dálnici u Velkého Meziříčí...*[42] I přes poměrně vysokou frekvenci lemmat, vytvářejí kolokaci s námi zkoumaným lemmatem *automobil*.

Mezi kolokáty s frekvencí nižší nebo rovno čtyřem, které míra vygenerovala, patří kolokát *průjezd* na čtvrté pozici s frekvencí dva a hodnotou 17.770 ve významu např.: *...drahou tzv. Bránu nebo jiné zařízení, které zaznamená průjezd automobilu...*[43] Na páté pozici se nachází kolokát *luxusní s* frekvencí tři a hodnotou

[37] [lemma="automobil"]
[38] Key word in context (klíčové slovo v kontextu) v tomto případě je to pro nás slovo automobil.
[39] Všechna lemmata, která tvoří kolokaci s KWIC, jsme se pro přehlednost rozhodli označovat jen jako kolokát, ta která netvoří buď, jako lemma, popřípadě jinak.
[40] Dufka, J., Vytápění netradičními zdroji tepla, 2003, str. 364.
[41] Hospodářské noviny, 2.6.2005, str. 46.
[42] Respekt, č. 3/2005, str. 31.
[43] Respekt, č 3/2005, str. 57.

17.114 ve významu např.: *...nejsmrtelnější pokles křivky zůstatkové hodnoty vykazují luxusní automobily...*[44] Na šesté pozici kolokát *koupě* s frekvencí dva a hodnotou 16.981 ve významu například *...jeho směna na eura a koupě vysněného automobilu...*[45] Na desáté pozici kolokát *vybavit* s frekvencí dvě a hodnotou 15.981 v kontextu ve větě například *...Automobil je vybaven palubní jednotkou a u vjezdů a výjezdů...*[46]

Od druhé poloviny řady převládají synsémantika a interpunkční znaménka s vyšší frekvencí, např.: na pozici třicet dva předložka *z* s frekvencí sedm a hodnotou 11.399. Čtyřicátou čtvrtou pozici obsadila spojka *a* s hodnotou 10.154 a frekvencí dvacet šest, to je nejvyšší frekvence lemmatu celé vygenerované řady. Předložka *na*, kterou míra vygenerovala na třicátou osmou pozici, má frekvenci dvanáct a hodnotu 10.002.

Posledních dvanáct vygenerovaných lemmat řady má frekvenci dvě a jen jedno z těchto dvanácti lemmat je schopno tvořit kolokaci s námi zkoumaným lemmatem *automobil*. Kolokát *majitel* s frekvencí dvě a hodnotou 15.163 ve významu např.: *...by zřejmě bylo chápáno jako omezení svobody občana, majitele automobilu...*[47]

3.1.2 T-score

Míra t-score udala řadu o celkových šedesáti dvou pozicích, kde od první poloviny převládají synsémantika s vyšší frekvencí. Do dvacáté čtvrté pozice mají frekvenci vyšší nebo rovno čtyřem. První pozici obsadilo interpunkční znaménko . s frekvencí třicet a hodnotou 5.470. Druhou pozici také obsadilo interpunkční znaménko , s frekvencí dvacet osm a hodnotou 5.281. Na třetí pozici míra vygenerovala spojku *a* s frekvencí dvacet čtyři a hodnotou 5.095. Na prvních pozicích s nejvyšší hodnotou jsou tedy synsémantika, která nejsou schopná vytvářet kolokace.

I přes to, že vygenerovaná lemmata řady mají vysokou frekvenci, udala míra t-score sedm solidních kolokátů. Na páté pozici se nachází kolokát *osobní* s frekvencí třináct

[44] Hospodářské noviny, 2.6.2005, str. 22.
[45] Hospodářské noviny, 2.6.2005, str. 20.
[46] Respekt, č.3/2005, str. 55.
[47] Respekt, č. 3/2005, str. 17.

a hodnotou 3.606 ve významu např.: *..současně musíme vylepšit a modernizovat normy pro spotřebu paliva osobních automobilů...* [48] Na deváté pozici kolokát *nákladní* s frekvencí osm a hodnotou 2.828. Kolokát na šestnácté pozici *užitkový* s frekvencí pět s hodnotou 2.236. Na pozici sedmnáct, se stejnou frekvencí pět a hodnotou 2.236 vygenerovala míra kolokát *nový* ve významu např.: *...poptávka po nových automobilech se pohybuje v cyklech trvajících kolem pěti a osmi let..*[49] Na pozici devatenácté kolokát *nákup* s frekvencí čtyři a hodnotou 2.000. Na pozici dvacáté páté kolokát *luxusní* s frekvencí tři a hodnotou 1.732. Pozici dvacet šest obsadil kolokát *financování* se stejnou frekvencí a hodnotou, ve významu např.: *...zejména v segmentu financování automobilů je místo klasického finančního leasingu k dispozici hned několik produktů…* [50] A posledním z kolokátu je lemma *provoz* na dvacáté osmé pozici, také s totožnou frekvencí a hodnotou, ve významu např.: *...jaký je význam zavádění katalyzátorů, jestliže produkce i provoz automobilů nadále astronomicky rostou...*[51]

V druhé polovině řady převládají autosémantika s nižší t-score hodnotou a frekvencí, ta je u lemmat menší nebo rovna třem.

3.1.3 Log likelihood

Míra log likelihood vygenerovala výsledek, který je velice podobný výsledku řady t-score a od čtyřicáté deváté pozice jsou výsledky totožné.

Nejvyšší rozdíl mezi mírami je u lemmatu *že* s frekvencí tři, které obsadilo třicátou šestou pozici u míry t-score s hodnotou 1.729 a u míry log likelihood čtyřicátou osmou pozici s hodnotou 31.720. Rozdíl mezi lemmatem je dvanáct pozic. Dále interpunkční znaménko *“* s frekvencí čtyři obsadilo u míry t-score dvacátou čtvrtou pozici s hodnotou 1.995, u míry log likelihood pozici třicet pět s hodnotou 39.454. Rozdíl je tedy jedenáct pozic. Lemma *ten* s frekvencí čtyři obsadilo u míry t-score dvacátou třetí pozici s hodnotou 1.995, u míry log likelihood třicátou třetí pozici s hodnotou 40.500.

[48] Cílek, V., Kašík, Martin, Nejistý plamen, 2007, str. 2789.
[49] Cílek, V., Kašík, Martin, Nejistý plamen, 2007, str. 188.
[50] Hospodářské noviny, 2. 6. 2005, str. 20.
[51] Britské listy, 12.5. 2005, str. 33.

Míra log likelihood se zachovala o něco víc uživatelsky příznivěji a lemmata, která nejsou vhodná pro vytváření kolokace, zařadila na nižší pozice oproti míře t-score.

3.1.4 LogDice

Míra logDice vygenerovala řadu, kde první polovina udává poměrně solidní výsledek, do třicáté první pozice převažují lemmata s nižším počtem výskytů, převážně je frekvence menší nebo rovna čtyřem.

Výjimkou jsou čtyři lemmata na první pozici kolokát *nákladní* s frekvencí osm a hodnotou 11.093. Na druhé pozici kolokát *osobní* s frekvencí třináct a hodnotou 10.712. Na třetí pozici kolokát *užitkový* s frekvencí pět a hodnotou 10.368 a kolokát *nový* na pozici dvacet sedm s frekvencí pět a hodnotou 7.104. I přes jejich vysokou frekvenci, utváří kolokaci s lemmatem *automobil*.

Ve druhé polovině řady jsou převážně synsémantika s vysokou frekvencí, např. lemma *pro* na třicáté druhé pozici s frekvencí osm a hodnotou 6.128. Na třicáté páté pozici spojka *i* s frekvencí devět a hodnotou 5.885 nebo spojka *a* s frekvencí dvacet šest a hodnotou 4.920. Míra ve druhé polovině řady vygenerovala převážně lemmata, která nejsou schopná tvořit kolokaci.

3.1.5 Minimální citlivost

Minimální citlivost vygenerovala podobný výsledek jako míra logDice, rozdíl mezi výslednými řadami je vždy jen pár pozic.

Nejvyšší rozdíl mezi lemmaty je u lemmatu *automobil* s frekvencí dvě, které obsadilo u míry logDice patnáctou pozici s hodnotou 8.371, u minimální citlivosti osmou pozici s hodnotou 0.020. Rozdíl mezi lemmaty je sedm pozic, u tohoto lemmatu se však nejedná o kolokát, jde jen o náhodný spolu výskyt ve větě, např.: *...nejstrmější pokles křivky zůstatkové hodnoty vykazují luxusní automobily, automobily se silnými benzinovými motory...*[52] Lemma *nadstandard* s frekvencí dva obsadilo u míry logDice sedmou pozici s hodnotou 9.342, u minimální citlivosti

[52] Hospodářské noviny, 2. 6. 2005, str. 22.

obsadila čtrnáctou pozici s hodnotou 0.020, stejně jako u předchozího lemmatu je rozdíl sedm pozic a ani v tomto případě se nejedná o kolokaci, jen o náhodný souvýskyt slov, např. ve větě *...velké a menší firmy stále ještě považují osobní i užitkový automobil za jakýsi nadstandard...* [53] Lemma *koupě* s frekvencí dvě obsadilo u míry logDice devátou pozici s hodnotou 9.117, u minimální citlivosti patnáctou pozici s hodnotou 0.020 a lemma *financování* s frekvencí tři obsadilo u míry logDice desátou pozici s hodnotou 9.077, u minimální citlivosti pátou pozici s hodnotou 0.030. U ostatních lemmat je rozdíl v řadách vygenerovaný měrami menší nebo rovno dvěma a od čtyřicáté deváté pozice jsou řady totožné.

3.2 Lemma pán

Lemmatu *pán* jsme zmenšit rádius KWICu na mínus dva nalevo a dva napravo.

3.2.1 MI-score

Míra MI-score vygenerovala solidní řadu, která má celkově sto devadesát jedna pozic, kde lemmata mají nižší frekvenci, až na šest pozic je frekvence lemmat nižší nebo rovna pěti. Do padesáté pozice není frekvence lemmat řazena sestupně, ale zpřeházeně, a od padesáté první pozice je řada řazená sestupně od frekvence pět do počtu výskytů lemmat rovné dvěma.

Jak bylo řečeno výše, šest lemmat má vyšší frekvenci, jsou to následující lemmata: kolokát *urozený* s frekvencí osm a hodnotou 17.374 na druhé pozici v kontextu, např.: *...jako z udělání potkal hned u lesa urozeného pána, co mu patřil ten les a všechno kolem...* [54]. *N*a dvanácté pozici lemma *dáma* s frekvencí devět a hodnotou 15.059. Lemma *Agáta* s frekvencí šest a hodnotou 14.822. Lemma na dvacáté sedmé pozici *Jakub* s frekvencí dvacet sedm a hodnotou 13.847. Interpunkční znaménko *:* na třicáté třetí pozici s hodnotou 13.39 má nejvyšší frekvenci dvě stě dvacet sedm z celé řady. Lemma *ano* s frekvencí deset a hodnotou 13.179 obsadilo třicátou osmou pozici, další pozici třicátou devátou obsadil kolokát *bůh* s frekvencí sedm a hodnotou 13.146 v kontextu např.: *...prazákladními rysy lidské povahy, s během přírody, s Pánem Bohem nebo s kosmickým děním...* [55]

[53] Hospodářské noviny, 2. 6. 2005, str. 2.
[54] Hospodářské noviny, 2. 6. 2005, str. 2.
[55] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 161.

Z těchto šesti lemmat jsou jen dvě lemmata, která tvoří kolokaci s lemmatem *pán*.

Mezi kolokáty s nižší frekvencí, které se vešly do padesáté pozice, patří kolokáty: na první pozici kolokát *farár* s frekvencí dva a hodnotou 17.696 ve významu např.: *...Hyba se stala, pan farár, hyba!...*[56] Na třetí pozici kolokát *šedovlasý* s frekvencí dva a hodnotou 17.111 ve významu *...vmísil s do rozhovoru (mimořádně laskavým tónem) šedovlasý pán...*[57] *N*a deváté pozici kolokát *árijský* s frekvencí dva a hodnotou 16.374 ve významu např.: *...zbytek se měl stát negramotnými otroky pracujícími pro své árijské pány nově dobytých územích n východě...*[58]

Míra vygenerovala řadu uživatelsky příznivou, z celkové řady obsadily první místa autosémantika, která jsou schopná tvořit solidní kolokace.

3.2.2 T-score

Míra t-score vygenerovala na první pozice převážně synsémantika a interpunkční znaménka s vysokou frekvencí. První pozici obsadilo interpunkční znaménko . s frekvencí 256 a hodnotou 15.922. Druhou pozici obsadilo interpunkční znaménko *:* s frekvencí 229 a hodnotou 15.131. Na třetí pozici míra vygenerovala také interpunkční znaménko , s frekvencí 119 a hodnotou 10.881. Nižší frekvence lemmat začíná až od padesáté čtvrté pozice, kde lemma *aby* má frekvenci čtyři a hodnotu 1.996, od této pozice je frekvence lemmat řazena sestupně od počtu výskytu čtyři do dvou.

Do padesáté čtvrté pozice jsou jen dvě lemmata schopná tvořit kolokaci s lemmatem *pán*. Na sedmnácté pozici kolokát *starý* s frekvencí sedmnáct a hodnotou 4.122, v kontextu ve větě např.: *...jistě by starému pánovi chyběla, s židovským a německým živlem ztratilo město bezprostřední...*[59] Kolokát *urozený* na třicáté deváté pozici s frekvencí osm a hodnotou 2.828.

Míra se nezachovala uživatelsky vhodně, pokud uživatel hledá zajímavé kolokace,

[56] Vaculík, L., Morčata, 2004, str. 17.
[57] Kundera, M., Směšné lásky, str. 389.
[58] Britské listy, 12. 5. 2005, str. 13.
[59] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 16.

musí se probírat obrovským množstvím synsémantik a interpunkčních znamének, které míra vygenerovala na první pozice. Úměrně pak platí, že čím nižší frekvence, tím vyšší počet kolokátů v řadě, nejvíc jich tak najdeme až na konci řady, a to nepovažujeme za uživatelsky příznivé.

3.2.3 Log likelihood

Mira log likelihood vygenerovala řadu, která je velmi podobná řadě míry t-score. Rozdíly ve výsledcích jsou jen do padesáté pozice z celkových sto devadesáti dvou, a to většinou jen do deseti pozic.

Vyšší rozdíl než deset pozic mají jen tři lemmata, kolokát *urozený* s frekvencí osm obsadilo u míry log likelihood devatenáctou pozici s hodnotou 186.359, u míry t-score pozici třicátou devátou s hodnotou 2.996, rozdíl mezi vygenerovanými řadami je tedy dvacet pozic. Lemma *dáma* s frekvencí devět obsadilo u míry log likelihood pozici dvacet dva s hodnotou 171.554, u míry t-score obsadila pozici třicet šest s hodnotou 3.000. Předložku *v* míra log likelihood vygenerovala na pozici čtyřicátou s hodnotou 102.150, míra t-score o sedmnáct pozic výše na dvacátou třetí pozici.

Odlišně se míry zachovaly jen u lemmatu *melancholicky* s frekvencí čtyři, které míra loglikelihood vygenerovala na pozici třicáté čtvrté s hodnotou 88.595. Míra t-score lemma *melancholický* neuvedla, naopak vygenerovala kolokát *znát* na padesáté pozici s frekvencí pět a hodnotou 2.235 ve významu např.: *...znám oba pány už docela obstojně a držím palce panu doktorovi Kreiskému...*[60]

3.2.4 LogDice

Řada vygenerovaná mírou logDice udala do čtvrté pozice lemmata s velmi vysokou frekvencí. Na první pozici interpunkční znaménko *:* s frekvencí 229 a hodnotou 10.529. Na druhé pozici lemma *Jakub* s frekvencí dvacet sedm a hodnotou 10.529. Třetí pozici obsadilo interpunkční znaménko *!* s frekvencí čtyřicet sedm a hodnotou 9.314. Čtvrtou pozici obsadilo interpunkční znaménko *?* s frekvencí padesát šest a hodnotou 9.128.

[60] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 157.

Oproti mírám t-score a log likelihood, jež také vygenerovala lemmata s vysokou frekvencí na první pozice, míra logDice udala převážně autosémantika a čtyři z nich do dvacáté první pozice považujeme za kolokáty. Na páté pozici kolokát *náš* s frekvencí dvacet šest a hodnotou 8.949 ve významu např.:. *...to bylo stanoveno v okamžiku, kdy náš pán nás stvořil...*[61] Na sedmé pozici kolokát *urozený* s frekvencí osm a hodnotou 8.813. Kolokát *starý,* který míra vygenerovala na desátou pozici s frekvencí sedmnáct a hodnotou 8.663. Třináctou pozici obsadil kolokát bůh s frekvencí sedm a hodnotou 8.282 v kontextu např.: *...a byla to osoba příliš dobrá, aby ji Pán Bůh potrestal tím, že by se dožila komunismu...*[62] A na sedmnácté pozici lemma *můj* s frekvencí dvanáct a hodnotou 8.146.

Od dvacáté druhé pozice převažují lemmata s frekvencí nižší nebo rovno pěti. Úměrně s nižším počtem lemmat s vysokou frekvencí přibývá počet kolokátů v řadě, do padesáté pozice jsme vyhodnotili šest lemmat jako kolokáty. Jsou to lemmata *milý* na pozici dvacet šest s frekvencí čtyři a hodnotou 7.539 v kontextu např.: *...znovu uviděla své tělo a to ji vzrušilo. Milí pánové, Alžběta měla před vámi všemi jednu výhodu...* [63] Kolokát *Kristus* s frekvencí tři a hodnotou 7.376 vygenerovala míra logDice na pozici třicet tři ve větě např.: *...ošidným nápisem : UKŘIŽOVÁNÍ KRISTA PÁNA BUDIŽ CHLOUBA NAŠE...*[64] Na pozici čtyřicáté udala míra logDice lemma *mladý* s frekvencí pět a hodnotou 7.275 v kontextu např.: *...to byla také doba, kdy páni profesoři posílali mladé pány v parném červnu domů pro vázanky...*[65] Pozici čtyřicet sedm obsadilo lemma *společný* s frekvencí tři a hodnotou 7.056 ve významu např.: *...mladý básník, který jednoho dne přišel za naším společným pánem?...* A lemma *bohatý* na čtyřicáté deváté pozici s frekvencí tři a hodnotou 7.021 ve významu např.: *...šaty a dobrého koně, zkrátka žil s jako bohatý pán...*[66]

3.2.5 Minimální citlivost

Minimální citlivost a míra logDice vygenerovaly do padesáté pozice podobné řady, které se liší vždy jen o pár pozic a od padesáté první pozice jsou výsledky totožné.

[61] Kundera, M., Jakub a jeho Pán, 1992, str. 28.
[62] Škvorecký, J., Ráda zpívám z not a jiné eseje, 2004, str. 98.
[63] Kundera, M., Směšné lásky, 1991, str. 193.
[64] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 158.
[65] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 93.
[66] Vladislav, J., Pohádky paní Meluzíny, 1999, str. 121.

Opačně se míry zachovaly jen u osmi pozic, míra logDice vygenerovala odlišná lemmata oproti minimální citlivosti, na třicáté pozici lemma *Saint-Ouenovi* s frekvencí tři a hodnotou 7.408. Kolokát *Kristus* na třicáté třetí pozici s frekvencí tři. Kolokát *vážený* na pozici třicet pět s frekvencí tři. Lemma *práh* na pozici třicet šest a frekvencí tři a hodnotou 7.771. Interpunkční znaménko . obsadilo pozici třicet devět s frekvencí dvě stě padesát čtyři a hodnotou 7.335. Kolokát *mlčet* vygenerovala míra na pozici čtyřicet jedna s frekvencí tři a hodnotou 7.270 ve významu např.: *...Utopit se? (Pán mlčí) Vrazit si meč do prsou?...*[67] Na pozici čtyřicáté sedmé míra vygenerovala kolokát *společný* s frekvencí tři. Kolokát *bohatý,* obsadil pozici čtyřicet devět s frekvencí tři.

Těchto osm lemmat minimální citlivost nevygenerovala, naopak udala odlišná lemmata. Na pozici třicet sedm kolokát *sám* s frekvencí čtyři a hodnotou 0.005 v kontextu např.: *...brána paláce se otevřela a na prahu stál sám pán domu...*[68] Lemma *rád* na pozici třicet devět s frekvencí tři a hodnotou 0.005, které netvoří kolokaci s námi zkoumaným lemmatem *pán*, jde jen o náhodný souvýskyt lemmat ve větě. Stejně tak netvoří kolokaci lemma *vést* na čtyřicáté pozici s frekvencí tři a hodnotou 0.005. Kolokát *zámek* na pozici čtyřicet tři s frekvencí tři a hodnotou 0.005 v kontextu např.: *...unesl ji rovnou z oken královského zámku mocný obr, pán tohoto zámku...*[69] Kolokát *rozhodl* na pozici čtyřicet čtyři s frekvencí tři a hodnotou 0.005, v kontextu např.: *...HOSTINSKÁ: váš pán rozhodne...*[70] Pozici čtyřicet pět obsadil kolokát *mnoho* s frekvencí tři a hodnotou 0.005 v kontextu např.: *...je mnoho autorit a mnoho emírů a poddaní podléhají mnoha pánům...*[71] Kolokát *ministr* obsadil čtyřicátou sedmou pozici s frekvencí tři a hodnotou 0.005 v kontextu např.: *...A bude to na vás, páni minitři...*[72] A lemma *dívat* pozici čtyřicet osm s frekvencí tři a hodnotou 0.005, toto lemma ale nevytváří kolokaci.

3.3 Lemma šťastný

Pro lemma šťastný jsme nastavili vyšší okruh slov na mínus tři nalevo od KWICu

[67] Kundera, M., Jakub a jeho pán, 1992, str. 12.
[68] Vladislav, J., Pohádky paní Meluzíny, 1999, str. 91.
[69] Tamtéž, str. 25.
[70] Kundera, M., Jakub a jeho pán, 1992, str. 35.
[71] Cílek, V., Kašík, M., Nejistý plamen, 2007, str. 375.
[72] Trefulka, J., Bláznova čítanka – fejetony, str. 107.

a tři napravo.

3.3.1 MI-score

Míra vygenerovala solidní výsledek, do dvacáté páté pozice z celkových šedesáti dvou převažují autosémantika a z nich deset považujeme za kolokáty.

Na prvním místě kolokát *manželství* s frekvencí čtyři a hodnotou 17.573 v kontextu např.: *...největší neštěstí, jaké vás může potkat, je šťastné manželství...*[73] Druhou pozici obsadil kolokát *náhoda* s frekvencí dva a hodnotou 16.229 v kontextu např.: *...je prvním klíčem, na který hmátl, byla to šťastná náhoda...*[74] Na třetí pozici míra vygenerovala kolokát *učinit* s frekvencí dva a hodnotou 17.573 v kontextu např.: *...času, než mu markýza nakonec podlehla a učinila šťastným...*[75] Kolokát *osud*, který míra MI-score vygenerovala na čtvrtou pozici s frekvencí dva a hodnotou 14.688 v kontextu *...tolik bylo napsáno o ní, jejím hlase i nešťastném osudu...*[76] Pozici devátou obsadil kolokát *konec* s frekvencí tři a hodnotou 13.204 ve významu např.: *...se v tomto tak vzácném příběhu se šťastným koncem zázraku nedočkal...*[77] Kolokát *chvíle*, který obsadil třináctou pozici s frekvencí dva a hodnotou 12.644 ve větě např.: *...překoná strmý sráz a verše budu znít, dcery mých šťastných chvil...*[78] Na sedmnáctou pozici míra vygenerovala kolokát *ruka* s frekvencí dva a hodnotou 15.525 ve významu např.: *...a jak se později ukázalo, měl při výběru místa šťastnou ruku...*[79] Kolokát *celý* obsadil dvacátou pozici s frekvencí čtyři a hodnotou 12.203 *...starý král byl celý šťastný, že José přinesl, co si přála jeho milovaná...* [80] S frekvencí dva obsadil kolokát *žena* dvacátou první pozici s hodnotou 11.778 v kontextu např.: *...Šťastný to muž, šťastná to žena. Inu, ta Praha...*[81] A na dvacátou druhou pozici vygenerovala míra kolokát *člověk* s frekvencí čtyři a hodnotou 11.681 ve větě např.*: ...že i v největší bídě, může být člověk šťastný...*[82]

[73] Kundera, Milan, Směšné lásky, 1991, str. 173.
[74] Vaculík, Ludvík, Morčata, 2004, str. 178.
[75] Kundera, Milan, Jakub a jeho pán, 1992, str. 25.
[76] Reflex, č. 51/2005 str. 32.
[77] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 48.
[78] Škvorecký, J. Ráda zpívám z not a jiné eseje, 2004. str. 154.
[79] Rubín, J. (ed.), Přírodní klenoty České republiky, 2006, str. 479.
[80] Vladislav, J., Pohádky paní Meluzíny, 1999, str. 261.
[81] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 281.
[82] Aktuálně.cz, 31. 3. 2006. str. 18.

Od druhé poloviny řady převažují synsémantika a interpunkční znaménka s vysokou frekvencí, a to až do počtu výskytů čtyřicet pět. Na úkor synsémantik a interpunkčních znamének ubyl i počet kolokátů. Ve druhé polovině řady míra vygenerovala jen dva kolokáty. Na padesáté čtvrté pozici kolokát *chvíle* s frekvencí dva a hodnotou 10.704 v kontextu např.: *...smuteční hosté se bez obřadu rozcházejí, byla to šťastná chvíle...*[83] A na pozici padesát sedm vygenerovala míra MI-score kolokát *velice* s frekvencí dva a hodnotou 13.573 v kontextu např.: *...proto jsem velice šťastná z mé zprávy...*[84]

Míra vygenerovala řadu, která je uživatelsky příznivá, zajímavé kolokace generuje již od prvních pozic, naopak lemmata, která jsou synsémantická, udává na nižší pozice řady, uživatel tak dostane zajímavé kolokace na první místa.

3.3.2 T-score

Míra t-score vygenerovala řadu o šedesáti dvou pozicích, kde v první polovině převažují synsémantika s vysokou frekvencí, nejvyšší počet výskytů má spojka *a* na první pozici s frekvencí čtyřicet šest a hodnotou 4.791.

Do první poloviny řady míra vygenerovala jen čtyři kolokáty. Na patnácté pozici kolokát *manželství* s frekvencí čtyři a hodnotou 2.000. O pozici níž, na pozici patnáct, míra udala kolokát *král* se stejnou frekvencí čtyři a hodnotou 2.000. Kolokát *člověk* obsadil devatenáctou pozici s frekvencí čtyři a hodnotou 1.999 a kolokát *konec* na dvacáté páté pozici s frekvencí tři a hodnotou 1.732.

Ve druhé polovině řady ubylo množství synsémantik a vzrostl počet kolokátů na osm. Na pozici třicet dva míra vygenerovala kolokát *rok* s frekvencí tři a hodnotou 1.731. Pozici třicet čtyři obsadil kolokát *náhoda* s frekvencí dva a hodnotou 1.414. Kolokát *učinit* se stejnou frekvencí dva a hodnotou 1.414 obsadil pozici třicet pět. Se stejnou frekvencí dva a hodnotou 1.414 obsadil kolokát *osud* pozici třicet šest. Na pozici třicet osm dosadila míra t-score kolokát *velice* s frekvencí dva a hodnotou 1.414 v kontextu např.: *...proto jsem velice šťastná, že mně se podařilo postavit*

[83] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 73.
[84] Reflex, č. 51/2005 str. 21.

si velkou kariéru... [85] Kolokát *ruka* obsadil pozici čtyřicet čtyři s frekvencí dva a hodnotou 1.414. S frekvencí dva a hodnotou 1.414 obsadil kolokát *žena* pozici čtyřicet čtyři. Na padesátou čtvrtou pozici vygenerovala míra t-score kolokát *chvíle* s frekvencí dva a také hodnotou 1.414.

Z celé řady je patrné, že zajímavé kolokace mají frekvenci dva nebo tři, ve chvíli, kdy je frekvence vyšší, je větší pravděpodobnost, že se jedná jen o náhodný souvýskyt slov, nebo o interpunkční znaménko v okolí KWICu. V celé naší řadě jsme vyhodnotili jen tři kolokace, které mají vyšší frekvenci než dva a zároveň je považujeme za zajímavé. Jsou to již výše zmíněná lemmata *manželství*, *král* a *člověk*. T-score se nezachovala příznivě pro uživatele, kolokace se převážně vyskytly až ve druhé polovině vygenerované řady.

3.3.3 Log likelihood

Míra log likelihood vygenerovala podobné výsledky jako míra t-score, do sedmé a od třicáté deváté pozice jsou řady totožné. Rozdíly od osmé do třicáté osmé pozice jsou vždy jen pár pozic.

Nejvyšší rozdíl má předložka *na* s frekvencí pět, ta obsadila u míry t-score pozici čtrnáctou s hodnotou 2.231, u míry log likelihood pozici dvacet jedna s hodnotou 51.168. Na patnácté pozici míra t-score vygenerovala kolokát *manželství* s frekvencí čtyři, u míry log lokelihood obsadila osmou pozici s hodnotou 90.264. Pro obě lemmata to je rozdíl pouhých sedm pozic.

Od třicáté deváté pozice se obě míry chovají totožně, všechna lemmata od této pozice níže mají frekvenci dva, takže kolokace, které můžou být neotřelé a uživatelsky zajímavé, uživatel najde až ke konci řady. To hodnotíme jako uživatelsky nepříznivé.

3.3.4 LogDice

Míra logDice vygenerovala solidní řadu o celkových šedesáti dvou pozicích. V první polovině převažují autosémantika s nižší frekvencí menší nebo rovno čtyřem.

[85] Reflex, č. 51/2005, str. 37.

Dvanáct lemmat z první poloviny vygenerované řady považujeme za kolokáty.

Na první pozici je kolokát *manželstvi* s frekvencí čtyři a hodnotou 0.042. Stejnou frekvenci čtyři a hodnotu 0.021 mají kolokáty: na druhé pozici kolokát *osud*, na pozici třetí kolokát *náhoda* a na čtvrté pozici kolokát *učinit.* Kolokát *král* míra vygenerovala na sedmu pozici s frekvencí čtyři a hodnotou 0.010. Pozici osmou obsadil kolokát *velice* s frekvencí dva a hodnotou 0.010. Kolokát *konec* s frekvencí tři a hodnotou 0.007 obsadil devátou pozici. Na šestnáctou pozici vygenerovala míra kolokát *způsob* s frekvencí dva a hodnotou 0.005 ve významu např.: *...zapadnout do party nebo si tímto nešťastným způsobem léčit bolavou duši...* [86] Pozici sedmnáctou obsadil kolokát *ruka* s frekvencí dva a hodnotou 0.005. Kolokát *celý* vygenerovala míra na dvacátou pozici s frekvencí čtyři a hodnotou 0.004. Na dvacátou první pozici vygenerovala míra kolokát *žena* s frekvencí dva a hodnotou 0.003. A kolokát *člověk*, který míra vygenerovala na dvacátou druhou pozici s frekvencí čtyři a hodnotou 0.003.

Ve druhé polovině vygenerované řady převažují synsémantika a interpunkční znaménka a žádná z těchto pozic v druhé polovině řady nevytváří kolokaci k našemu lemmatu *šťastný*. Míra se zachovala uživatelsky velice příznivě, když všechny kolokace vygenerovala do první poloviny řady.

3.3.5 Minimální citlivost

Minimální citlivost nám sice vygenerovala také slušnou výslednou řadu, ale od předchozí míry MI-score se liší jen minimálně, nejvyšší rozdíl je u kolokátu *osud* s frekvencí dva, který míra logDice vygenerovala na pátou pozici s hodnotou 0.021, u minimální citlivosti obsadil kolokát druhou pozici s hodnotou 8.415.

3.4 Lemma politický

Pro lemma *politický* jsme snížili okruh lemmat okolo našeho KWICu na mínus dva nalevo a dva napravo, počet frekvence kolokátu v kontextu a korpusu jsme nechali stejný, tedy na dvou výskytech.

[86] Hospodářské noviny, 2. 6. 2005.

3.4.1 MI-score

Míra nám vygenerovala opravdu solidní výsledky, a to i ke konci řady řada má celkově sto třicet devět pozic.

Na první místa míra vygenerovala lemmata s frekvencí dva nebo tři, tedy s frekvencí nižší. Vyšší frekvenci najdeme až od čtyřicáté první pozice, kde se nachází kolokát *strana* s hodnotou 13.633 ve významu např.: *...ODS chce zrušit konkursy, situací se začaly zabývat i politické strany...*[87]. má frekvenci dvacet.

Prvních deset pozic vygenerovaných mírou považujeme všechny za kolokáty. Na první pozici míra vygenerovala kolokát *zákulisí* s frekvencí dva a hodnotou 17.548 v kontextu např.: *...že jen sebral dlouhou řadu historek a drbů z politického zákulisí...* [88] Druhou pozici obsadil kolokát *seskupení* s frekvencí tři s hodnotou 17.455 ve větě např.: *...zlé je, že toto politické seskupení si umínilo vnucovat nám své duchovní tradice a způsoby...*[89]

Kolokát *namířený* obsadil třetí pozici s frekvencí dva a hodnotou 17.285 v kontextu např.: *...což se pak bere jako proti Straně namířená politická činnost...*[90] Pozici čtvrtou obsadil kolokát *intrika* s frekvencí dva a hodnotou 17.062 ve větě např.: *...historik a pletichář namočený do příliš mnoha politických intrik...* [91] S frekvencí dva a hodnotou 16.870 obsadil kolokát *thriller* v kontextu např.: *...má nyní vzniknout politický thriller...*[92] Na šestou pozici vygenerovala míra kolokát *rival* s frekvencí dva a hodnotou 16.700 ve větě např.: *...peníze protékaly těmi správnými spřátelenými firmami, a nikoli kanály politických rivalů...*[93] Sedmou pozici obsadil kolokát *spektrum* s frekvencí tři a hodnotou 16.548 ve významu např.: *...nejpodstatnější změnou v německém politickém spektru je však úspěch nově vytvořeného stranického útvaru Die Linke...*[94] Kolokát *fráze* s frekvencí dva obsadil osmou pozici s hodnotou 16.548 v kontextu např.: *...a přece jenom nepochybně vlastenecký a minimálně obložený politickou frází...* [95] Na devátou pozici míra

[87] Britské listy, 12. 5. 2005.
[88] Respekt, č. 3/2005.
[89] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 163.
[90] Škvorecký, J., Ráda zpívám z not a jiné eseje, 2004, str. 13.
[91] Cílek, V., Kašík, M., Nejistý plamen, 2007, str. 201.
[92] Hospodářské noviny, 2. 6. 2005
[93] Týden, č. 48/2005
[94] Britské listy, 19. 9. 2005.
[95] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 88.

vygenerovala kolokát *nátlak* s frekvencí dva a hodnotou 16.410 ve větě např.: *...neznámé panoptikum kalkulů a pletich, všudypřítomné korupce, politického nátlaku a soudu...*[96] Desátou pozici obsadil kolokát *provokace* s frekvencí dva a hodnotou 16.548 v kontextu např.: *...dotčená dívka označila spolužákův smích za politickou provokaci...*[97]

Řada vygenerovaná mírou MI-score udala z celkových sto třiceti devíti pozic čtyřicet sedm kolokátů, které utvářejí kolokace s námi zkoumaným lemmatem *politický*. Většina z těchto kolokátů se navíc nachází v první polovině vygenerované řady. Míra se tedy zachovala uživatelsky velice příznivě.

3.4.2 T-score

Míra t-score oproti předchozí míře nevygenerovala tak obstojné výsledky, do dvacáté pozice vygenerovala převážně interpunkční znaménka a synsémantika s vysokou frekvencí, vyšší nebo rovno pěti. První pozici s hodnotou 8.586 s frekvencí sedmdesát čtyři obsadilo interpunkční znaménko , druhou pozici obsadila spojka *a* s frekvencí padesát devět a hodnotou sedm. Interpunkční znaménko , obsadilo třetí pozici s frekvencí padesát tři a hodnotou 7.265.

Přes vysokou frekvenci lemmat do dvacáté pozice vygenerovala míra tři, která jsou součástí kolokace s naším lemmatem *politický*. Jsou to kolokáty *strana* na pozici šesté s frekvencí dvacet a hodnotou 4.472 Na patnácté pozici míra vygenerovala kolokát *činnost* s frekvencí pět a hodnotou 2.236 A kolokát *situace*, který obsadil šestnácté místo s frekvencí pět a hodnotou 2.236 ve významu např.: *...po skutečně levicové politické straně, která by hájila zájmy zaměstnanců a nebohatých...*[98]

S klesající frekvencí lemmat roste počet kolokátů v řadě. Od dvacáté první pozice míra vygenerovala lemmata s frekvencí rovno nebo menší čtyřem a od pozice padesáté osmé mají všechna vygenerovaná lemmata frekvenci rovnou dvěma. Mezi pozicemi dvacet jedna a padesáté dva vygenerovala míra celkově devět kolokátů.

[96] Respekt, č. 3/2005.
[97] Kundera, M., Směšné lásky, str. 308.
[98] Britské listy, 19. 9. 2005, str. 20.

Na pozici dvacáté první *kariéra* ve významu např.: *...roce 1968 dokončila studium práv a souběžně s politickou kariérou působila od roku 1970 jako právnička...*[99] Pozici dvacet dva obsadil kolokát *kritika* ve významu např.: *...filmového scénáře, v níž režim zřejmě snese i hodně politické kritiky, ale jako non-fiction...*[100] Kolokát *scéna o*bsadil pozici dvacet tři v kontextu např.: *...žádný intelekt nezabrání tomu, aby vznikla politická scéna se vším, co nás většinou tak irituje...*[101] A na dvacátou pátou pozici míra vygenerovala kolokát *právo* v kontextu např.: *...případě jejich klientů byla porušena Mezinárodní úmluva o občanských politických právech...*[102] S frekvencí tři a hodnotou 2.000 míra udala na třicáté páté pozici kolokát *seskupení* ve větě např.: *...zlé je, že toto politické seskupení si umínilo vnucovat nám své duchovní tradice a způsoby...*[103] Kolokát *spektrum* obsadil pozici třicet šest. Na pozici třicátou sedmou dosadila míra kolokát *shoda* ve významu např.: *...vláda: vytvoří, co nejširší politickou shodu ohledně přijetí nejlepší formy důchodové reformy...*[104] Pozici čtyřicátou obsadil kolokát *reprezentace* v kontextu např.: *...když někteří ekonomové a většina špiček politické reprezentace vypadají, že si to snad myslí...*[105] A na pozici čtyřicátou čtvrtou míra vygenerovala kolokát *tlak v* kontextu např.: *...nebude-li rozložena vnějšími politickými tlaky...*[106]

Míra t-score vygenerovala stejný počet kolokátů jako míra MI-score, nevýhoda této míry však je v tom, že kolokáty se nenachází převážně v první polovině, ale prostupují celou řadou, a uživatel tak musí manuálně třídit všechny výsledky řady.

3.4.3 Log likelihood

Stejně jako u předchozích lemmat je míra t-score daleko více ovlivněna frekvencí n-gramů, zatímco míra Log likelihood výsledky neřadí striktně sestupně podle frekvence.

[99] Hospodářské noviny, 2. 6. 2005.
[100] Škvorecký, J., Ráda zpívám z not a jiné eseje, 2004. str. 13.
[101] Reflex, č. 51/2005.
[102] Hospodářské noviny, 2. 6. 2005.
[103] Trefulka, J., Bláznova čítanka – fejetony, nakladatelství Atlantis, 1998, str. 178.
[104] Britské listy, 12. 5. 2005.
[105] Britské listy, 12. 5. 2005.
[106] Cílek, V., Kašík, M., Nejistý plamen, 2007, str. 161.

Největší rozdíly jsou pak od dvacáté pozice do pozice padesáté a od padesáté první pozice jsou výsledky totožné. Nejvyšší rozdíly míry vygenerovaly mezi interpunkčním znaménkem *“* s frekvencí pět, to míra t-score vygenerovala na dvacátou pozici s hodnotou 2.224, míra log likelihood na pozici čtyřicátou devátou s hodnotou 42.050. Částici *že* s frekvencí pět, kterou míra t-score vygenerovala na pozici devatenáctou s hodnotou 2.230, však míra log likelihood dosadila až na pozici třicátou šestou s hodnotou 48.511. Interpunkční znaménko *(* míra t-score dosadila na osmnáctou pozici také s frekvencí pět a hodnotou 2.231, míra log likelihood na pozici třicátou čtvrtou s hodnotou 50.405.

Odlišné výsledky míry vygenerovaly v případě t-score na třicáté druhé pozici spojku *ale* s frekvencí čtyři a hodnotou 1.996. Předložka *k* obsadila třicátou třetí pozici s frekvencí čtyři a hodnotou 1.995. Lemma *mít* míra vygenerovala na třicátou čtvrtou pozici s frekvencí čtyři a hodnotou 1.994. Pozici čtyřicet sedm obsadilo lemma *některý* s frekvencí tři a hodnotou 1.731. Lemma *český* s frekvencí tři a hodnotou 1.731 obsadilo pozici čtyřicátou osmou. Na čtyřicátou devátou pozici míra vygenerovala lemma *celý* s frekvencí tři a hodnotou 1.731. A na pozici padesáté vygenerovala míra lemma *také* s frekvencí tři a hodnotou 1.730. Žádné z těchto lemmat však nepovažujeme za kolokát.

Míra log likelihood naopak vygenerovala sedm lemmat, jež všechna považujeme za součást kolokace s námi zkoumaným lemmatem *politický*. Na třicáté osmé pozici míra vygenerovala kolokát *zákulisí* s frekvencí dva a hodnotou 45.604. Pozici čtyřicátou obsadil kolokát *namířený* s frekvencí dva a hodnotou 45.604. Na pozici čtyřicátou druhou uvedla míra kolokát *intrika* s frekvencí dva a hodnotou 43.958. Pozici čtyřicátou třetí obsadil kolokát *thriller* s frekvencí dva a hodnotou 43.336. Na pozici čtyřicátou pátou míra vygenerovala kolokát *fráze* s frekvencí dva a hodnotou 42.326. Kolokát *nátlak* obsadil pozici padesátou s frekvencí dva a hodnotou 41.903. Míra log likelihood vygenerovala uživatelsky daleko příznivější řadu oproti t-score.

3.4.4 LogDice

Míra LogDice sice upřednostňuje na první pozice lemmata s vyšší frekvencí,

ale ne do takové míry jako míra t-score, výsledkem jsou pak solidní kolokace. Do prvních padesáti pozic míra udala dvacet osm kolokátů, z nich osm žádná z předchozích měr neudala. Na dvanácté pozici kolokát *vězeň* s frekvencí tři a hodnotou 8.400 v kontextu např.: *...pomáhat nevládním organizacím, které podporují rodiny disidentů a jejich politickou činnost...*[107] Na třicáté pozici míra vygenerovala kolokát *elita* s frekvencí dva a hodnotou 7.907 ve větě např.: *...Holanďané nicméně mají pocit, že politické elity se o to nepokusily...* [108] Pozici třicátou první obsadil kolokát *dění* s frekvencí dva a hodnotou 7.886 v kontextu např.: *...určité části naší inteligence, jak se projevovaly na pozadí politického dění Československu a vyvrcholily u nás v krizi koncem...*[109] Na třicátou třetí pozici míra dosadila kolokát *soupeř* s frekvencí dva a hodnotou 7.871 ve větě např.: *...který se těší všeobecné úctě u veřejnosti i politických soupeřů...*[110] Kolokát *integrace* obsadil třicátou čtvrtou pozici s frekvencí dva a hodnotou 7.860 v kontextu např.: *...z tohoto hlediska nám může velmi pomoci rychlá faktická, tedy nejen politická integrace* v *E*vropské *unii...*[111] Na třicátou šestou pozici míra vygenerovala kolokát *orientace* s frekvencí dva a hodnotou 7.815 ve větě např.: *...útočníci údajně nijak nevyjádřili svou politickou orientaci...*[112] Pozici třicet devět obsadil kolokát *představitel* s frekvencí dva a hodnotou 7.795 v kontextu např.: *...bohužel někteří naši političtí představitelé jak na lokální, tak celostátní úrovni, si tyto hodnoty často neuvědomují ...*[113] A na pozici čtyřicátou druhou vygenerovala míra kolokát *osobnost* s frekvencí dva a hodnotou 7.738 ve větě např.: *...všechny tyto tahy ilustrují Foldynovu politickou osobnost...*[114]

Podobně jako u předchozích příkladů, i tady má míra logDice tendenci řadit výsledná lemmata sestupně podle jejich frekvence, oproti míře MI-score, která lemmata řadí nevázaně na jejich frekvenci. Výsledkem jsou pak velice solidní kolokace.

[107] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 228.
[108] Hospodářské noviny, 2. 6. 2005 str. 34.
[109] Škvorecký, J., Ráda zpívám z not a jiné eseje, 2004. str. 51.
[110] Respekt, č. 3/2005.
[111] Rubín, J., Přírodní klenoty České republiky, 2006, str. 57.
[112] Hospodářské noviny, 2. 6. 2005.
[113] Rubín, J., Přírodní klenoty České republiky, 2006, str. 55.
[114] Respekt, č. 3/2005.

3.4.5 Minimální citlivost

Na rozdíl od předchozího lemmatu se tentokrát míra zachovala do padesáté pozice rozdílně oproti míře logDice, od padesáté první pozice vygenerovaly míry stejné řady.

Do padesáté pozice se všechny výsledky minimální citlivosti a míry logDice liší, kromě první pozice, kam obě míry vygenerovaly kolokát *strana* s frekvencí dvacet a hodnotou pro míru logDice 9.345, hodnota u minimální citlivosti 0.027. Nejvyšší rozdíl je u kolokátu *projev,* který u minimální citlivosti obsadil pozici dvacet osm s hodnotou 0.008 a u míry logDice až pozici čtyřicátou osmou s hodnotou 7.565, tedy dvacet pozic rozdíl.

Obě míry vygenerovaly devět odlišných lemmat. Pro míru logDice jsou to lemmata, jež všechna mají frekvenci dva, na osmnáctou pozici míra vygenerovala lemma *zákulisní* s hodnotou 7.983. Pozici dvacet šest obsadil kolokát *politický* s hodnotou 7.940. Kolokát *dění* míra vygenerovala na pozici třicet jedna s hodnotou 7.886. Pozici třicet tři obsadil kolokát *soupeř* s hodnotou 7.871. Pozici třicet osm obsadilo lemma *korupce* s hodnotou 7.805. Na čtyřicáté první pozici udala míra lemma *vůle* s hodnotou 7.738. Na pozici čtyřicátou první míra vygenerovala lemma *premiér* s hodnotou 7.660.

Minimální citlivost vygenerovala na dvacáté pozici kolokát *moc* s frekvencí tři a hodnotou 0.010 ve významu např.: *...ať už z titulu politické moci nebo soustředěných finančních prostředků...*[115] Pozici dvacet čtyři obsadilo lemma *směr* s frekvencí dva a hodnotou 0.008. Na dvacátou pátou pozici míra udala kolokát *hranice* s frekvencí dva a hodnotou 0.008 v kontextu např.: *...oblastí probíhala jako tzv. "železná opona" politická hranice mezi Východem a Západem...*[116] Kolokát *boj* obsadil třicátou třetí pozici s frekvencí dva a hodnotou 0.008 v kontextu např.: *...standard života musíme vybrat od občanů odpovídající sumu peněz a politický boj se vede o to...*[117] Na pozici třicáté šesté lemma *sociální* s frekvencí dva a hodnotu 0.008, na třicáté sedmé udala míra lemma *rámec* s frekvencí dva

[115] Trefulka, J., Bláznova čítanka – fejetony, 1998, str. 170.
[116] Rubín, J., Přírodní klenoty České republiky, 2006, str. 16.
[117] Britské listy, 12. 5. 2005.

a hodnotou 0.008, pozici třicátou devátou obsadilo lemma *přítel* s frekvencí dva a hodnotou 0.008l, na čtyřicátou čtvrtou pozici udala míra lemma *vzniknout* a na čtyřicátou sedmou lemma *německý*.

Obě míry se tedy zachovaly podobně, udaly sice devět různých lemmat v řadě, ze všech devíti lemmat však jen tři tvořily kolokáty u každé z měr.

3.5 Lemma místy

Pro lemma *místy* jsme parametry upravili, zvýšili jsme okruh okolo KWICu na mínus tři nalevo a tři napravo. Důvodem byl nízký výskyt lemmatu v našem subkorpusu – lemma má v celém našem subkorpusu počet výskytu devadesát šest.

3.5.1 MI-score

Míra MI-score vygenerovala výsledek o padesáti sedmi pozicích, kde autosémantika mají převahu jen do první poloviny řady.

Z celkových sedmdesáti pěti pozic míra vygenerovala šest kolokátů. Kolokát *káňonovitý* na páté pozici s frekvencí tři a hodnotou 17.514 ve významu například *...pod soutěskou je údolí místy káňonovité a na svazích vystupují stěny a skalní hřebínky...* [118] Kolokát *ráz* s frekvencí tři a hodnotou 17.085 na osmé pozici ve významu například *...na nejdolnějším toku poblíž soutoku s Vltavou sevřeným skalnatým údolím místy soutěskovitého rázu...* [119] Na třináctou pozici míra vygenerovala kolokát *tmavý* s frekvencí dva a hodnotou 16.344 v kontextu např.: *...s nepravidelnými (zubatými) okraji, místy tmavší (případně světlejší), zvětšuje se a zesiluje...* [120] Pozici čtrnáct obsadil kolokát *setkávat* s frekvencí dva a hodnotou 16.344 ve významu např.: *...při toulkách přírodou v našich horách a pahorkatinách se místy setkáváme s přirozenými výchozy skalních hornin...* [121] Kolokát *vystupovat* míra vygenerovala na pozici devatenáct s frekvencí dva a hodnotou 15.699 ve významu např.: *...místy vystupují i malé skalky bez vegetace...* [122]

[118] Rubín J. (ed.), Přírodní klenoty České republiky, 2006. str. 204.
[119] Tamtéž, str. 163.
[120] Hospodářské noviny, 2. 6. 2005.
[121] Rubín, J. (ed.), Přírodní klenoty České republiky, 2006, str. 263.
[122] Tamtéž, str. 366.

Na pozici dvacet čtyři míra udala kolokát *vyskytova*t s frekvencí tři a hodnotou 14.645 v kontextu např.: *...místy se vyskytují mrazové sruby, izolované skály, balvanové proudy...* [123] Pozici třicet devět obsadil kolokát *velký* s frekvencí tři a hodnotou 11.170 ve významu např.: *...při výlevu na mořské dno a výše pyroklastické uloženiny, místy s většími kusy diabasových láv...*[124]

Od třicáté první pozice převažují především synsémantika, a výsledkem pak jsou lemmata, jež nejsou schopna tvořit kolokace.

3.5.2 T-score

Míra t-score vygenerovala do desáté pozice interpunkční znaménka a synsémantika s poměrně vysokou frekvencí vzhledem k nízkému počtu výskytů lemmatu *místy* v našem subkorpusu.

Interpunkční znaménko , na první pozici s frekvencí třicet sedm a hodnotou 6.076. Na desáté pozici dosadila míra lemma *se* s frekvencí šest a hodnotou 2.444.
Naopak od jedenácté pozice míra vygenerovala autosémantika, z nichž šest považujeme za součást kolokace s lemmatem *místy*.

Na pozici patnáct vygenerovala míra kolokát *káňonovitý* s frekvencí tři a hodnotou 1.732. Kolokát *ráz* s frekvencí tři a hodnotou 1.732 vygenerovala míra na pozici šestnáct. Kolokát *vyskytovat* na dvacáté pozici s frekvencí tři a hodnotou 1.732. Na pozici dvacet tři vygenerovala míra kolokát *velký* s frekvencí tři a hodnotou 1.731. Pozici třicet pět obsadil kolokát *setkávat* s frekvencí dva a hodnotou 1.414. Na pozici čtyřicet šest vygenerovala míra kolokát *vzniknout* s frekvencí dva a hodnotou 1.416 ve významu např.: *...s hukotem se valí přes balvany, v nichž místy vznikly dokonalé tvary obřích hrnců...*125

Míra t-score neudala tak solidní výsledek jako míra MI-score. Zatímco v první polovině řady převažují synsémantika, kolokáty nacházíme až od druhé poloviny

[123] Tamtéž, str. 98.
[124] Tamtéž, str. 75.
[125] Tamtéž, str. 208.

řady.

3.5.3 Log likelihood

Míra log likelihood udala podobný výsledek jako míra t-score, první dvě pozice jsou stejné a od třicáté sedmé pozice jsou výsledky totožné, a to až do poslední padesáté sedmé pozice.

Rozdílné výsledky jsou tedy jen u lemmat od třetí do třicáté šesté pozice a maximální rozdíl mezi lemmaty činí dvanáct pozic, a to u kolokátů *velký* s frekvencí tři, to obsadilo u míry t-score dvacátou třetí pozici s hodnotou 1.731, u míry log likelihood třicátou pátou pozici s hodnotou 40.584. Stejný rozdíl je i u předložky *od* s frekvencí tři, ta obsadila u míry t-score dvacátou čtvrtou pozici s hodnotou 1.731, u míry log likelihood pozici třicet šest s hodnotou 40.309.

U většiny lemmat je ale rozdíl daleko nižší okolo čtyř pozic, například u kolokátu *setkávat* s frekvencí dva obsadilo u míry t-score třicátou pátou pozici s hodnotou 1.414, u míry log likelihood třicátou první pozici s hodnotou 41.474.

3.5.4 LogDice

Míra logDice vygenerovala do první poloviny solidní výsledek. Z celkových padesáti sedmi pozic do třicáté první převažují autosémantika s frekvencí, která není vyšší než tři, od třicáté druhé pozice naopak převažují synsémantika s vyšší frekvencí než tři, a to až dvacet.

Míra vygenerovala celkově osm kolokátů. První pozici obsadil kolokát *káňonovitý* s frekvencí tři a hodnotou 9.927, druhou pozici lemma *ráz* s frekvencí tři a hodnotou 9.804. Na třináctou pozici míra vygenerovala kolokát *tmavý* s frekvencí dva a hodnotou 9.167. Kolokát *setkávat* obsadil patnáctou pozici s frekvencí dva a hodnotou také 9.167. Na pozici dvacet vygenerovala míra kolokát *vystupovat* s frekvencí dva a hodnotou 8.696. Pozici dvacátou druhou obsadil kolokát *vyskytovat* s frekvencí tři a hodnotou 8.535. Kolokát *vzniknout* vygenerovala míra na pozici třicátou s frekvencí dva a hodnotou 7.674. Kolokát *velký* obsadil pozici třicet devět

s frekvencí tři a hodnotou 5.490.

Jak bylo řečeno výše, od druhé poloviny řady převažují synsémantika a interpunkční znaménka s vyšší frekvencí. Například spojka *i* na pozici třicáté sedmé s frekvencí osm a hodnotou 5.722. Interpunkční znaménko *)* na pozici třicáté osmé s frekvenci devět a hodnotou 5.641. Spojka *a* na čtyřicáté šesté pozici má frekvencí dvacet a interpunkční znaménko *,* na pozici čtyřicet osm má dokonce počet výskytů třicet sedm.

3.5.5 Minimální citlivost

Minimální citlivost vygenerovala řadu, jež je od třicáté druhé pozice totožná s mírou logDice, do třicáté první pozice se výsledek liší, ale vždy jen o pár pozic.

Nejvyšší rozdíl je u lemmatu *melancholický* s frekvencí dva, to obsadilo u míry logDice třetí pozici s hodnotou 9.715, u minimální citlivosti obsadilo lemma šestnáctou pozici s hodnotou 0.027, rozdíl je tedy 13 pozic. Lemma *lutea* s frekvencí dva obsadilo u míry logDice čtvrtou pozici s hodnotou 9.715, minimální citlivost lemma vygenerovala na patnáctou pozici s hodnotou 0.027. Dále lemma *plavba* s frekvencí dva obsadilo u míry logDice osmou pozici, minimální citlivost lemma vygenerovala až na pozici osmnáctou. Lemma *koryto* s frekvencí tři obsadilo u míry logDice desátou pozici s hodnotou 9.425, u minimální citlivosti druhou pozici s hodnotou 0.040. Ve zbylých případech je rozdíl mezi pozicemi vygenerovaných řad do šesti.

3.6 Lemma rychle

Pro lemma *rychle* jsme snížili okruh okolo KWICu na mínus dva nalevo a dva napravo z důvodu vyšší frekvence lemmatu v našem subkorpusu, ta se rovná 227. Minimální frekvenci kolokátu v korpusu a kontextu jsme ponechali na dvou.

3.6.1 MI-score

Výsledná řada udala velice solidní výsledek, lemmata mají spíše nižší frekvenci, a to většinou frekvenci dva, nebo tři. Výsledná lemmata jsou převážně autosémantika.

Do prvních padesáti pozic míra vygenerovala osmnáct kolokátů. Na první pozici míra udala kolokát *opadávat* s frekvencí dva a hodnotou 19.032, například v kontextu *...přítoky do řeky Moravy z východu velmi rychle opadávají...*[126] Na druhé pozici s frekvencí dva vygenerovala míra kolokát *zblednout*, například v kontextu *...zdálo se mi, že rychle zbledl...* [127] Kolokát *vyběhnout* obsadil šestou pozici s frekvencí dva a hodnotou 15.944 v kontextu např.: *...už jsem musela Pavla chytat, aby nespadl, tak rychle vyběhl...* [128] Na pozici sedmé míra vygenerovala kolokát *stoupat* s frekvencí pět a hodnotou 15.376 ve větě např.: *...voda stoupá rychle, zaslechl jsem zde i pár němců...*[129]

Na pozici sedmé míra vygenerovala kolokát *jíst* s frekvencí čtyři a hodnotou 15.250 v kontextu např.: *...jez rychleji mládenče...*[130] Kolokát *běžet* obsadil osmou pozici s frekvencí sedm a hodnotou 15.058 ve větě např.: *...že by mě tělo tolik nebolelo, kdybych běžel rychleji...* [131] Kolokát *reagovat* na pozici desáté s výskytem tři ve významu například *...že bude obviněn z podvodu, a tak rychle reaguje a obviní z téhož podvodu druhou stranu...* 132 Pozici jedenáctou obsadil kolokát *rostoucí* s frekvencí čtyři a hodnotou 14.882 v kontextu např.: *...k významným rychle rostoucím dřevinám s řadí vrba, olše, akát...* [133] Na dvanácté pozici míra vygenerovala kolokát *poměrně* s frekvencí osm a hodnotou 14.424 ve větě např.: *...tak i aplikace pro něj se neustále a poměrně rychle se vyvíjí...*[134] Čtrnáctou pozici obsadil kolokát *otočit* s frekvencí dva a hodnotou 14.174 v kontextu např.: *...případný náznak překvapení ale ruský celek rychle odmítl a stejně rychle otočil skóre...*[135] Kolokát *klesat* vygenerovala míra na pozici patnáctou s frekvencí dva a hodnotou 14.101 ve větě např.: *...produkce rychle klesala...*[136] Pozici šestnáctou obsadil kolokát *přejít* s frekvencí dva a hodnotou 14.055 v kontextu např.: *...to vás*

[126] Aktuálně.cz, 31. 3. 2006.
[127] Reflex, č. 51/2005.
[128] Vaculík, L., Morčata, 2004, str. 374
[129] Aktuálně.cz, 31. 3. 2006.
[130] Vladislav, J., Pohádky paní Meluzíny, 1999, str. 370.
[131] Blesk, 4. 4. 2005.
[132] Cílek, V. – Kašík, M., Nejistý plamen, 2007, str. 111.
[133] Dufka, J., Vytápění netradičními zdroji tepla, 2003, str. 198.
[134] Tamtéž, str. 231.
[135] Aktuálně.cz, 31. 3. 2006.
[136] Cílek, V. – Kašík, M., Nejistý plamen, 2007, str. 67.

pak rychle přejde smích...[137] Kolokát *zavést* obsadil sedmnáctou pozici s frekvencí dva a hodnotou 13.842 v kontextu např.: *...akademickou a státní sférou a na tuzemský trh co nejrychleji zavést směrnice Evropské unie...*[138] Na pozici devatenáctou míra vygenerovala kolokát *ztratit* s frekvencí čtyři a hodnotou 14.033 ve významu např.: *...ho sice varovně koutku mysli, ale pak se rychle ztratil a široko daleko zůstali jen ona a muž...*[139] Pozici dvacátou druhou obsadil kolokát *růst* s frekvencí šest a hodnotou 13.229 ve větě např.: *...musel přizpůsobit unii výši daní a zároveň v té době rychle rostl objem úvěrů...*[140] Kolokát *velice* vygenerovala míra na pozici dvacet tři s frekvencí tři a hodnotou 12.916 v kontextu např.: *...a tím i ceny dopravy potravin to však mohou velice rychle změnit...*[141] Kolokát *velmi* obsadil pozici dvacet čtyři s frekvencí devět a hodnotou 12.851 ve významu např.: *...růstu vývozu je česká ekonomika v krajním případě schopna velmi rychle zaplatit...*[142] Na pozici dvacet pět vygenerovala míra kolokát *připravit* s frekvencí dva a hodnotou 12.851 v kontextu např.: *...musíme se rychle připravit...*[143]

Míra se zachovala uživatelsky velice příznivě. Do první poloviny řady vygenerovala osmnáct kolokátů, druhá polovina sice také tvoří převážně autosémantika, většinu z nich však nepovažujeme za součást kolokace s naším lemmatem *rychle*.

3.6.2 T-score

Míra t-score vygenerovala na první pozice převážně synsémantika s vysokou frekvencí, nejvyšší frekvenci má interpunkční znaménko , s frekvencí sedmdesát čtyři na první pozici s hodnotou 8.588, do páté pozice pak mají lemmata frekvenci vyšší nebo rovnou dvacet šest, dál do dvacáté první pozice mají lemmata vyšší frekvenci nebo rovno pěti.

Přesto míra do první poloviny řady vygenerovala deset kolokátů. Na osmé pozici kolokát *velmi* s frekvencí devět a hodnotou 3.000. Pozici desátou obsadil kolokát

[137] Reflex, č. 51/2005.
[138] Hospodářské noviny, 2. 6. 2005.
[139] Kundera, M., Směšné lásky, 1991, str. 196.
[140] Hospodářské noviny, 2. 6. 2005.
[141] Cílek, V. – Kašík, M., Nejistý plamen, 2007, str. 225.
[142] Aktuálně.cz, 31. 3. 2006.
[143] Cílek, V. – Kašík, M., Nejistý plamen, 2007, str. 261.

poměrně s frekvencí osm a hodnotou 2.828. Kolokát *běžet* míra vygenerovala na pozici třinácté s frekvencí sedm a hodnotou 2.646. Pozici dvacátou třetí obsadil kolokát *jíst* s frekvencí čtyři a hodnotou 2.000. Kolokát *stoupat* míra vygenerovala na pozici dvacet čtyři s frekvencí čtyři a hodnotou 2.000. Na pozici dvacet pět míra udala kolokát *rostoucí* s frekvencí čtyři a hodnotou 2.000. Pozici dvacátou šestou obsadil kolokát *začít* s frekvencí čtyři a hodnotou 1.999, v kontextu např.: *...na přítoku se to ale začalo rychle zvedat a vody přišlo naráz daleko víc...*[144] Kolokát *reagovat* obsadil pozici třicet čtyři s frekvencí tři a hodnotou 1.732 ve významu např.: *...který má v rezervě velké čerpací kapacity je schopný rychle reagovat...*[145] Na pozici třicet pět míra vygenerovala kolokát *ztratit* s frekvencí tři a hodnotou 1.732. Pozici třicet sedm obsadil kolokát *velice* s frekvencí tři a hodnotou 1.732.

Za uživatelsky nepříznivé považujeme to, že míra udala skoro o polovinu méně kolokátů v první polovině řady oproti míře MI-score.

3.6.3 Log likelihood

Vygenerovaný výsledek udal do šesté pozice stejný výsledek jako míra t-score, od sedmé pozice do padesáté jsou výsledky různé, avšak liší se vždy jen o pár pozic a od padesáté první pozice jsou řady opět totožné.

Naprosto odlišný výsledek udávají míry u pěti lemmat, pro míru t-score to je předložka *v* na pozici třicáté druhé s frekvencí čtyři a hodnotou 1.981. Na čtyřicáté sedmé pozici lemma *velký* s frekvencí tři a hodnotou 1.730. Lemma *oni*, to obsadilo pozici čtyřicet osm s frekvencí tři a hodnotou 1.729. Lemma *který* obsadilo čtyřicátou devátou pozici s frekvencí tři a hodnotou 1.725. Předložku *z* míra vygenerovala na padesátou pozici s frekvencí tři a hodnotou 1.724. Žádné z těchto lemmat míra log likelihood neuvedla, ale žádné z lemmat není součástí kolokace.

Naopak míra log likelihood vygenerovala na dvacáté sedmé pozici kolokát *opadávat* s frekvencí dva a hodnotou 52.785, třicátou první pozici obsadil kolokát *zblednout*

[144] Aktuálně.cz, 31. 3. 2006.
[145] Cílek, V. – Kašík, M., Nejistý plamen, 2007, str. 88.

s frekvencí dva a hodnotou 48.966. Na pozici čtyřicátou vygenerovala míra lemma *účinně* s frekvencí dva a hodnotou 41.005. Pozici čtyřicátou první obsadil kolokát *vyběhnout* s frekvencí dva a hodnotou 40.470. Lemma *utéci* vygenerovala míra na pozici čtyřicátou devátou s frekvencí dva a hodnotou 35.910.

Tři z pěti lemmat, která vygenerovala míra loglikelihood tvoří kolokaci s lemmatem *rychle*. Míra log likelihood se v tomto případě zachovala daleko uživatelsky příznivěji.

3.6.4 LogDice

Vygenerovaná řada míry logDice udala na první pozice lemmata s vyšším počtem výskytů v našem subkorpusu. Do deváté pozice je rozmezí frekvence od čtyř do devíti počtů výskytů, frekvence lemmat na prvních pozicích řady tedy není tak vysoká jako u měr t-score a log likelihood, ale je vyšší než u míry MI-score.

Výsledkem jsou pak v první polovině řady převážně autosémantika, z nichž devatenáct tvoří kolokaci s lemmatem *rychle*. Na první pozici míra vygenerovala kolokát *běžet* s frekvencí sedm a hodnotou 9.411. Druhou pozici obsadil kolokát *poměrně* s frekvencí osm a hodnotou 9.279. Na čtvrtou pozici vygenerovala míra kolokát *jíst* s frekvencí čtyři a hodnotou 8.860. Pátou pozici obsadil kolokát *stoupat* s frekvencí čtyři a hodnotou 8.820. Šestou pozici obsadil kolokát *rostoucí* s frekvencí čtyři a hodnotou 8.781. Kolokát *reagovat* dosadila míra na osmou pozici s frekvencí tři a hodnotou 8.466. Na devátou pozici vygenerovala míra kolokát *velmi* s frekvencí devět a hodnotou 8.389. Jedenáctou pozici obsadil kolokát *opadavát* s frekvencí dva a hodnotou 8.161. Na pozici dvanáctou vygenerovala míra kolokát *zblednout* s frekvencí dva a hodnotou 8.155. Třináctou pozici obsadil kolokát *ztratit* s frekvencí tři a hodnotou 8.113. Na pozici šestnáct vygenerovala míra kolokát *vyběhnout* s frekvencí dva a hodnotou 8.069. Osmnáctou pozici obsadil kolokát *otočit* s frekvencí dva a hodnotou 7.845. Pozici devatenáct obsadil kolokát *klesat* s frekvencí dva a hodnotou 7.830. Dvacátou pozici obsadil kolokát *přejít* s frekvencí dva a hodnotou 7.820. Na dvacátou první pozici vygenerovala míra kolokát *velice* s frekvencí dva a hodnotou 7.820. Na pozici dvacet dva udala míra kolokát *zavést* s frekvencí dva a hodnotou 7.771. Na pozici dvacátou osmou míra udala kolokát

připravit s frekvencí dva a hodnotou 7.349. Čtyřicátou sedmou pozici obsadil kolokát *najít* s frekvencí dva a hodnotou 6.658 v kontextu např.: *...Paroubkův výrok našel rychle uplatnění...*[146]

Míra udala v první polovině řady devatenáct kolokátů, to považujeme za výsledek, který je uživatelsky příznivý.

3.6.5 Minimální citlivost

Minimální citlivost vygenerovala do padesáté pozice poměrně rozdílný výsledek oproti míře logDice.

Nejvyšší rozdíly zdaly pozice: kolokát *opadávat* s frekvencí dva, který u míry logDice obsadilo jedenáctou pozici s hodnotou 8.161, minimální citlivost jej vygenerovala na dvacátou sedmou pozici s hodnotou 0.009. Lemma *změnit* s frekvencí dva obsadilo u míry logDice třicátou šestou pozici, zatímco u minimální citlivosti sedmnáctou pozici. Kolokát *zblednout* s frekvencí dva obsadil u míry logDice dvanáctou pozici s hodnotou 8.155 a u minimální citlivosti třicátou třetí pozici s hodnotou 0.009.

Míry pro lemma *rychle* udaly poměrně rozlišné výsledky, totožná lemmata obsadila různé pozice, výsledkem jsou pak u obou měr převážně autosémantika, ale seřazená různě. Zatímco minimální citlivost u většiny případů lemmata řadí sestupně dle frekvence lemmat, míra logDice se nedrží tak striktně jejich frekvence. Odlišný výsledek udala míra logDice na čtyřicáté čtvrté pozici, kde vygenerovala interpunkční znaménko *!* s frekvencí sedm a hodnotou 6.789, minimální citlivost naopak udala kolokát *přijít* na pozici čtyřicet devět s frekvencí dva a hodnotou 0.004 v kontextu např.: *...letos ale voda do Znojma tak rychle nepřišla...*[147]

[146] Aktuálně.cz, 31. 3. 2006.

[147] Tamtéž.

Závěr

Ve své bakalářské práci jsem se zabývala základními statistickými mírami pro vytěžování jazykových dat a jejich využitím pro analýzu textů. V teoretické části jsem se pokusila nastínit základní pojmy, se kterými jsem dále pracovala. V praktické části jsem se testovala míry u lemmat s vysokou, nebo naopak nízkou frekvencí a popsala jejich chování.

Míra MI-score se ve všech zkoumaných případech zachovala uživatelsky velmi příznivě. U zkoumaných lemmat s vysokou i nízkou frekvencí vygenerovala většinu kolokátů do první poloviny řady. U lemmatu *politický* dokonce vygenerovala řada kolokáty, které obsadily prvních deset pozic. Míra je vhodná pro vyhledávání netypických kolokací pro daný korpus, tedy pro vyhledávání kolokátů s nízkou frekvencí, např. kolokaci *šedovlasý pán* míra vygenerovala k lemmatu *pán* již na třetí pozici s frekvencí dva. Míra může být použita na vyhledávání netypických slovních spojení jako frazémů a idiomů.

Míra t-score ve všech případech na první pozice vygenerovala interpunkční znaménka a synsémantika, která mají vysokou frekvenci. Výjimku tvořilo jen několik případů, např. kolokát *osobní* s frekvencí třináct obsadil pátou pozici v řadě vygenerované pro lemma *automobil*, v téže řadě se kolokát *nákladní* umístil na deváté pozici s frekvencí osm. Kolokát *starý* obsadil sedmnáctou pozici s frekvencí sedmnáct v řadě, kterou míra vygenerovala pro lemma *pán*. Oproti míře MI-score se míra t-score zachovala uživatelsky nepříznivě, obě míry sice vygenerovaly podobný počet kolokátů v řadách, ale míra t-score kolokáty vygenerovala převážně až do druhé poloviny řady, například u lemmatu *rychle* míra t-score vygenerovala skoro o polovinu méně kolokátů v první polovině řady. Uživatel tak musí výsledky pracně třídit. Míra je naopak vhodná na vyhledávání koligací a ustálených n-gramů.

Míra log likelihood se zachovala umírněněji oproti míře t-score. Například u lemmatu *automobil* udala na nižší pozice lemmata, jež nejsou součástí kolokace. Lemma *že* vygenerovala o dvanáct pozic níže a interpunkční znaménko *“* posunula níž dokonce o dvacet tři pozic. Daleko lépe oproti míře t-score se míra log likelihood zachovala u lemmat s vysokou frekvencí. V řadě pro lemma *politický* vygenerovala

sedm lemmat, která řada vygenerovaná kolokací t-score neuvádí. Všechna tato lemmata považujeme za kolokáty, míra t-score naopak uvedla sedm odlišných lemmat a ani jedno netvoří kolokaci s námi zkoumaným lemmatem *politický*. Podobně tomu je v řadě pro lemma *rychle*, míra log likelihood se zachovala uživatelsky příznivěji a uvedla pět odlišných lemmat oproti míře t-score. Z těchto pěti lemmat považujeme tři za kolokáty, naopak míra t-score uvedla pět lemmat, jež nejsou součástí kolokace. Míra je vhodná na vyhledávání ustálených slovních spojení a na práci s odbornou literaturou.

Míra logDice nepracuje s velikostí korpusu, ale s počtem výskytů lemmat v něm. Výsledky vygenerované řadou pak byly daleko příznivější, pokud námi zkoumané KWIC mělo nižší četnost výskytů v korpusu. Zatímco u lemmat s nižší frekvencí (např.: lemma *šťastný*) generuje míra už od prvních pozic kolokáty. První čtyři pozice vygenerované mírou logDice v řadě pro lemma *šťastný* obsadily kolokáty *manželství*, *osud*, *náhoda* a *učinit*. Naopak řady, které míra vygenerovala pro lemmata s vysokou frekvencí, udaly na první pozice převážně interpunkční znaménka a synsémantika, zajímavé kolokace najde uživatel až ve druhé třetině řady. Doporučujeme tedy používat míru logDice pro hledání neobvyklých kolokací v případě, že KWIC má frekvenci okolo sto padesáti.

Minimální citlivost se zachovala velmi podobně jako míra logDice. Pokud šlo o řady vygenerované pro KWIC s nízkým počtem výskytů, lišily se vždy jen o pár pozic. U KWICu s vyšší frekvencí vždy obě míry vygenerovaly řady, jež udaly několik odlišných lemmat, ale stejný počet kolokátů. U lemmatu *pán* obě míry vygenerovaly osm odlišných lemmat, z nichž je pět součástí kolokace. Podobná situace nastala u lemmatu *politický*, obě míry vygenerovaly pět různých lemmat, z nichž tři byly součástí kolokace. Minimální citlivost se zachovala uživatelsky příznivěji jen v jednom případě, a to u lemmatu *rychle*, kde vygenerovala kolokát *přijít*, na rozdíl od míry logDice, jež udala interpunkční znaménko *!* uživatel, který vyhledává neobvyklé kolokace k lemmatu s vyšší frekvencí, by měl využít obou měr, jelikož se vždy alespoň v několika případech zachovají odlišně. Míry logDice i minimální citlivost jsou vhodné pro vyhledávání frazémů a idiomů nebo na vyhledávání neobvyklých kolokací pro KWIC s nízkou frekvencí v korpusu.

Anotace

Autorka: Michaela Láníčková

Katedra:

Katedra bohemistiky, Filozofická fakulta Univerzity Palackého Olomouc

Název bakalářské práce:

Základní statistické metody pro vytěžování jazykových dat a jejich možnosti využití pro analýzu textů

Vedoucí práce: PhDr. Petr Pořízka, Ph. D.

Počet znaků: 82 286

Počet příloh: 2

Počet titulů použité literatury: 27

Klíčová slova:

korpus, ČNK, Český národní korpus, MI-score, t-score, log likelihood, logDice, minimální citlivost, asociační míry, kolokace.

Anotace:

Tato práce se zabývá popisem asociačních měr, které nabízí Český národní korpus. Vysvětluje, základní práci s mírami a technické parametry pro jejich využívání. Může sloužit jako základní manuál pro práci s mírami. První teoretická část je věnována definici kolokací, Českého národního korpusu a základním mírám, které ČNK nabízí. V praktické části se práce zabývá převážně zkoumáním měr u lemmat s různou frekvencí.

Abstract:

This bachelor thesis focuses on description of statistical methods provided by Czech national corpus. It explains a basic usage of methods and their technical parameters. It may be used as a basic manual for work with these methods. The first theoretical part deals with definitions of collocations, Czech national corpus and basic statistical methods provided by Czech national corpus. The practical part focuses mostly of studying described statistical methods in lemmas with different frequency.

Zdroje

Bibliografie

DUNNING, Ted. *Accurate methods for the statistics of surprise and coincidence* In*: Computational Linguistics*. 1993. [cit. 2018-02-19] dostupné: http://aclweb.org/anthology/J93-1003.

CVRČEK, Václav - RICHTEROVÁ, Olga (eds). [Internet]. Příručka ČNK; 2015 [cit. 2018-02-19]. Dostupné z: http://wiki.korpus.cz/.

ČERMÁK, František a Jan HOLUB. *Syntagmatika a paradigmatika českého slova*. 3. vyd. Praha: Karolinum, 2005. ISBN 80-246-0974-6.

ČERMÁK, František, Jana KLÍMOVÁ a Vladimír PETKEVIČ, ed. *Studie z korpusové lingvistiky*. In Praze: Karolinum, 2000. Acta Universitatis Carolinae. ISBN 80-7184-893-X.

ČERMÁK, František a Michal ŠULC, ed. *Kolokace*. Praha: NLN, Nakladatelství Lidové noviny, 2006. Studie z korpusové lingvistiky. ISBN 80-7106-863-2.

FIRTH, J. R. *Papers in linguistics, 1934-1951*. London: Oxford University Press, 1957. ISBN-10: 9027923108.

HLADKÁ, Zdeňka. a Petr. KARLÍK. *Čeština--univerzália a specifika: sborník konference ve Šlapanicích u Brna, 17.-18. 11. 1998*. Brno: Masarykova univerzita, 1999. ISBN 8021020253.

PEDERSEN, *Ted. Dependent Bigram Identification. In Department of Computer Science & Engineering*. 1994. . [cit. 2018-02-16] dostupné: https://www.aaai.org/Papers/AAAI/1998/AAAI98-193.pdf.

POŘÍZKA, Petr. *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*. V Olomouci: Vydavatelství Filozofické fakulty Univerzity Palackého, 2014. ISBN 978-80-87895-17-7.

RYCHLÝ, *Pavel. A Lexicographer-Friendly Association Score*. In: *RASLAN 2008*. 2. vyd. Brno, RASLAN 2008. Brno: Masarykova Univerzita, 2008. p. 6-9, 4 pp. ISBN 978-80-210-4741-9.

SINCLAIR, John McHardy. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991. Describing English language. ISBN 0-19-437144-1.

SKALIČKA, Vladimír. *Typ češtiny*. Praha: Slovanské nakladatelství, 1951. Slovanské jazykovědné příručky.

Použitá literatura a softwarový nástroj pro analýzu měr

BALATKA, Břetislav, RUBÍN, Josef, ed. *Přírodní klenoty České republiky*. Praha: Academia, 2006. ISBN 80-200-1377-6.

Britské listy [Internet]. 12. 5. 2005. ISSN 1213-1792 [cit. 2018-02-19]. Dostupné: https://legacy.blisty.cz.

CÍLEK, Václav a Martin KAŠÍK. *Nejistý plamen: průvodce ropným světem*. Praha: Dokořán, 2007. ISBN 978-80-7363-122-2.

DUFKA, Jaroslav. *Vytápění netradičními zdroji tepla: [biomasa - tepelná čerpadla - solární systémy]*. Praha: BEN - technická literatura, 2003. ISBN 80-7300-079-2.

Hospodářské Noviny IHNED [Internet]. 2. 6. 2005. 1996-2018 Economia, a.s., ISSN 1213-7693 [cit 2018-04-19]. Dostupné: https://archiv.ihned.cz.

KonText [Internet]. FF UK. Praha 2011. Dostupný z https://kontext.korpus.cz/ [cit. 2018-2-24].

KUNDERA, Milan. *Jakub a jeho pán: pocta Denisi Diderotovi*. Brno: Atlantis, 1992. ISBN 80-7108-032-2.

KUNDERA, Milan. *Směšné lásky: povídky*. Brno: Atlantis, 1991. ISBN 80-7108-027-6.

Reflex č. 51 [Internet]. 19.–25. 12. 2005 CZECH NEWS CENTER a.s. ISSN 1213-8991 [cit 2018-02-19]. Dostupné: http://www.reflex.cz/.

Respekt, č. 3 [Internet]. 16.–23. 1. 2005. Economia, a.s., ISSN 1801-1446. 7693 [cit. 2018-02-19]. Dostupné: https://www.respekt.cz.

ŠKVORECKÝ, Josef, ŠPIRIT, Michael, ed. *Ráda zpívám z not a jiné eseje*. Praha Ivo Železný, 2004. ISBN 80-237-3837-2.

TREFULKA, Jan. *Bláznova čítanka*. V Brně: Atlantis, 1998. ISBN 80-7108-169-8.

VACULÍK, Ludvík. *Morčata*. V Brně: Atlantis, 2004. ISBN 80-7108-248-1.

VLADISLAV, Jan, Jitka UHDEOVÁ a Jarmila VOJTOVÁ. *Pohádky paní Meluzíny*. Ilustroval Vlasta BARÁNKOVÁ. Brno: Atlantis, c1999. ISBN 80-7108-077-2.

Seznam obrázků

Resumé

This bachelor thesis focuses on basic statistical methods for the extraction of linguistic data and their use for text analysis. The thesis is divided into the theoretical and the practical part.

The theoretical part is focused on an explanation of general terminology, which were used further in the thesis. Especially terms as Czech national corpus, collocation and analysis methods. In the practical part lemmas of three different word classes with high or low frequency in the corpus were tested. I tried to described results of analytical methods of lemmas with high and low frequency and if they differ.

All results are summed up in the conclusion of this thesis. I conclude, that rate of MI-score is suitable for searching unusual collocations. Rate of t-score was almost in all cases user unfriendly and collocation were generated in the second half of the results and synsemantic predominantly in the first. The rate of loglikelihood is suitable for searching set phrases and idioms. The rate of logDice does operate with a size of corpus but with the number of lemmas occurrences. If the number of KWIC occurrences is around one hundred and fifty then it is suitable for searching unusual collocations. The results of minimal sensitivity were mostly similar as logDice but both of them should be used since they gave different results in several cases. The rate of LogDice and minimal sensitivity are suitable for searching phrases and idioms or for searching for unusual KWIC collocations with low frequency in corpus.

Příloha č. 1: Subkorpus

ID	Ator	Název	Rok vydání	Délka
Beletrie				
meluzina	Vladislav, Jan	Pohádky paní Meluzíny	1999	65876
ku_smlas	Kundera, Milan	Směšné lásky	1991	61622
eseje	Škvorecký, Josef	Ráda zpívám z not a jiné eseje	2004	61160
ku_jakub	Kundera, Milan	Jakub a jeho pán	1992	20904
trefulka	Trefulka, Jan	Bláznova čítanka - fejetony	1998	63174
morcata	Vaculík, Ludvík	Morčata	2004	53429
Počet slov				326165
Odborná literatura				
pneumol	Kolek, Vítězslav	Pneumologie pro magistry a bakaláře	2005	28807
sustavhs	Šlachtová, Hana	Suché stavby: konstrukce ze sádrokartonových a sádrovláknitých desek	2005	24864
linuxvb	Bednář, Vojtěch	Linux na firemním PC	2007	40890
vytapjd	Dufka, Jaroslav	Vytápění netradičními zdroji tepla	2003	20183

nizkoejt	Tywoniak, Jan	Nízkoenergetické domy: principy a příklady	2005	36603
nejplack	Cílek, Václav - Kašík, Martin	Nejistý plamen	2007	55169
klenotcr	Rubín, Josef (ed.)	Přírodní klenoty České republiky	2006	124009
Počet slov				330525
Publicistika				
bl050404		Blesk, 4. 4. 2005	2005	12720
tydn0548		Týden, č. 48/2005	2005	46421
bl051006		Blesk, 6. 10. 2005	2005	16327
by050512		Britské listy, 12. 5. 2005	2005	43700
ak060331		Aktuálně.cz, 31. 3. 2006	2005	39940
by050919		Britské listy, 19. 9. 2005	2005	8558
hn050602		Hospodářské noviny, 2. 6. 2005	2005	78394
refl0551		Reflex, č. 51/2005	2005	60115
resp0503		Respekt, č. 3/2005	2005	39532
Počet slov				345707
Slov celkem				1002397

Příloha č. 2: příkaz na vytvoření subkorpusu

<opus nazev="Pohádky paní Meluzíny" | nazev="Směšné lásky" | nazev="Ráda zpívám z not a jiné eseje" | nazev="Jakub a jeho pán" | nazev="Bláznova čítanka - fejetony" | nazev="Morčata" | nazev="Pneumologie pro magistry a bakaláře" | nazev="Suché stavby: konstrukce ze sádrokartonových a sádrovláknitých desek" | nazev="Linux na firemním PC" | nazev="Vytápění netradičními zdroji tepla" | nazev="Nízkoenergetické domy: principy a příklady" | nazev="Nejistý plamen" | nazev="Přírodní klenoty České republiky" | nazev="Blesk, 4. 4. 2005" | nazev="Týden, č. 48/2005" | nazev="Blesk, 6. 10. 2005" | nazev="Britské listy, 12. 5. 2005" | nazev="Aktuálně.cz, 31. 3. 2006" | nazev="Britské listy, 19. 9. 2005" | nazev="Hospodářské noviny, 2. 6. 2005" | nazev="Reflex, č. 51/2005" | nazev="Respekt, č. 3/2005" />