

Masarykova univerzita
Fakulta informatiky



Výpočet četnosti výskytů hesel SČFI v korpusu

Bakalářská práce

Jan Bušta

Brno, 2009

Prohlášení

Prohlašuji, že tato práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Jan Bušta

Vedoucí práce: Mgr. Pavel Rychlý, Ph.D.

Poděkování

Děkuji vedoucímu práce Mgr. Pavlu Rychlému, Ph.D. za jeho cenné rady a podněty při vypracovávání práce.

Shrnutí

Práce popisuje hledání idiomů ze Slovníku české frazeologie a idiomatiky, z částí Výrazy slovesné a Výrazy neslovesné. Idiomy jsou fráze, jejichž struktura je navzájem rozdílná. Tomu se muselo přizpůsobit i jejich vyhledávání, aby byl pokryt co možná největší počet výskytů idiomů a nebyla zanesena chyba. K vyhledávání slovesných a neslovesných frází byl vytvořen dotaz, pomocí kterého je dotazován korpus, zda obsahuje jednotlivé výrazy a v kolika výskytech. Dotaz má jinou (flexibilnější) podobu pro slovesná hesla, aby postihl volnější strukturu idiomů obsahujících sloveso, jinou pro neslovesné idiomy, které mají obecně pevnější vazby mezi jednotlivými slovy, než výrazy slovesné.

Abstract

This thesis describes searching the idioms from the dictionary of Czech idioms which includes verb-based and non-verb-based phrases. Idioms are phrases which structure is very different. This influences the searching of them to find as much idioms as possible but with keeping the error-rate. To find the verb-based as well as the non-verb-based phrases was created a general query. The application of query (querying the corpus) gives information if the corpus includes the idiom and how much occurrences it has. The query has different structure for each group of idioms because the verb-based are more flexible and should have many forms but the non-verb-based has more fixed bindings between words.

Klíčová slova

četnost idiomů, hledání slovních spojení, idiomatika, frazeologie, textový korpus, český jazyk, zpracování přirozeného jazyka

Keywords

frequency of idioms, multi-words expression searching, phraseology, text corpora, Czech language, natural language processing

Obsah

Úvod	9
1 Hesla SČFI	10
1.1 Změny v idiomech	11
1.1.1 Morfologická fixnost	11
1.1.2 Intersegmenty	11
1.1.3 Doplnění za neurčitá slova	12
1.1.4 Prohození slov	12
1.2 Výrazy neslovesné	12
1.3 Výrazy slovesné	13
1.4 Čištění dat	13
2 Korpus	14
3 Dotazování	15
3.1 Lemmatizace	16
3.2 Dotazovací jazyk	16
3.2.1 Operátor <code>meet</code>	17
3.2.2 Operátor <code>union</code>	18
3.3 Korpusový manager	18
4 Vytvoření dotazu	19
4.1 Vyhledávání slov	19
4.2 Obecný dotaz pro slovesné výrazy	19
4.3 Obecný dotaz pro neslovesné výrazy	20
4.4 Jednoslovná hesla	21
4.5 Vypouštění slov	21
4.6 Pokrytí	21
4.7 Přesnost	22
5 Četnost výrazů	23
5.1 Četnost slovesných výrazů	24
5.2 Četnost neslovesných výrazů	25

6	Další vývoj	27
	Závěr	29
	Literatura	30
	Přílohy	31
A	Četnosti slovesných idiomů	32
B	Četnosti neslovesných idiomů	33
C	Nejčtetnější slovesné idiomy	34
D	Nejčtetnější neslovesné idiomy	35
E	Četnost všech slovesných a neslovesných idiomů	36

Úvod

Při porozumění přirozenému jazyku je zásadním problémem význam. Přístup k jazyku se velice odlišný u člověka a u stroje. Přiřazení významu slovům je základním předpokladem toho, abychom si rozuměli. Cílem počítačového zpracování jazyka je snaha o totéž, o to, abychom dovedli komunikovat s počítačem pomocí přirozeného jazyka, avšak některé součásti jazyka jsou pro počítač zvlášť obtížně postihnutelné. Mezi ty rozhodně patří víceslovná spojení, mezi které idiomy patří.

Idiomy – dvou či víceslovná spojení, jejichž význam navíc není složen s významů jednotlivých slov, ale určuje ho až společný rozbor všech částí slovního spojení. Porušení principu kompozicionality[5] (neodvoditelnost významu z jednotlivých lexikálních složek) je vlastností idiomů a je otevřeným problémem, jak tuto abnormalitu idiomatických frází řešit.

Cesta, jak napomoci řešení, je vědět, jak časté jsou tyto fráze v běžně používaném jazyce, které jsou častější než jiné, které se v něm nevyskytují a do jaké míry je možné je zanedbávat či naopak na ně zaměřit pozornost.

Kapitola 1

Hesla SČFI

Hesla, která byla během práce zpracována, jsou výrazy ze Slovníku české frazeologie a idiomatiky (dále jen SČFI) z částí Výrazy slovesné a Výrazy neslovesné. Jako zdroj pro tuto práci sloužila elektronická verze výše zmíněných částí SČFI.

Kriterium pro výběr frází, které jsou v slovníku obsaženy, je „takové vymezení frazému, které za něj považuje každou ustálenou kombinaci alespoň dvou slovních forem téhož nebo různých slovních druhů, pro niž je příznačné to, že aspoň jeden z jejích komponentů se v daném významu a funkci kombinuje pouze s jedním komponentem dalším nebo s několika málo dalšími.“[1] Výběr frází je stejný pro slovesné i neslovesné výrazy. Vlastností idiomů je v tomto případě nějaká abnormalita, která je odlišuje od ostatních slovních spojení.

Míra abnormality či specifčnosti spojení slov se ale velice různí, ve slovníku jsou hesla jako *volný čas*, ale i třeba *udělat to do kalhot*. Zatímco v prvním případě je spojení velice volné a význam celého spojení lze odvodit z jednotlivých komponentních slov, druhé spojení je velice specifické a přináší přesah významů jeho komponentů. To, zda je slovní spojení idiomem, či nikoli, je i přes definici uvedenou výše individuální záležitostí.

SČFI lze považovat za synchronní slovník, výrazy, které se objevují jako hesla, nejsou v některých případech již plně srozumitelná a následně i jejich četnost v korpusu je velice malá – pokud se vůbec výraz v korpusu vyskytuje.¹ Na druhou stranu ve slovníku chybí některé fráze, které se postupem času staly běžnou součástí jazyka a zároveň splňují podmínky pro zařazení do slovníku.²

V některých případech je u idiomů uvedena i nespisovná podoba jeho slov. Je možné se setkat s výrazy typu *být/bejt*, *dobrý/dobrej*, *čich/čuch*, ale i s výrazy, kde je uvedena pouze nespisovná varianta (*bejt v cajku*), nebo jen spisovná podoba (*být na druhém břehu*). Nesjednocení tvarů v idiomech či

¹ *tuzexová slečna, bezpartijní bolševik*

² *dostat kapříka*

uvedení obou variant je motivováno nejspíš častějším používáním jednoho z tvarů (popřípadě obou). S ohledem na mluvenou češtinu, ve které se koncovky slov vyskytují převážně v nespisovné podobě, je nutno podotknout, že v podstatě každá fráze existuje také v nespisovné podobě, proto je uvádění alternativ spisovných a nespisovných slov zbytečné.

Slovník je určen k manuálnímu použití, kdy se autor opírá o částečné porozumění jazyku. To se projevuje především v naznačování alternativ v idiomech, kdy je stejné značení jak pro alternování jednotlivých slov, tak i pro víceslovná spojení. Rozhodnut, zda se jedná o předěl slov či slovních spojení je jednoduché pro čtenáře, který je na základě vlastního sémantického rozboru schopen přesně určit rozdíl, avšak počítačové zpracování zde naráží na problém. V počítačovém zpracování se ale o sémantickou intuici opřít nelze, protože není k dispozici.

1.1 Změny v idiomech

Jedním z největších problémů při vyhledávání idiomů v korpusu je jejich fixnost, resp. proměnlivost. Lexikální obsah idiomu lze v některých případech pozměňovat (morfologické změny – změna pádu, čísla nebo rodu u jmen; změna osoby, čísla, času či způsobu u sloves) nebo doplňovat o další lexikální jednotky (slova).

Ve slovníku jsou naznačeny u výkladových hesel slova, která jsou zaměnitelná, toto označení však není zcela úplné a nepostihuje všechna hesla slovníku.

Níže uvedené možnosti „nefixnosti“ idiomů se mohou navzájem kombinovat, a tím vzniká veliké množství možností, jak idiom vypadá. Chceme-li zachytit pokud možno všechny jeho formy, je třeba použít velice obecnou charakteristiku.

1.1.1 Morfologická fixnost

Některé idiomy obsahují slova, která jsou s výrazem svázána tak fixně, že rozdíl v pádě či čísle rozhoduje, zda se o idiom jedná, či nikoli.³ Údaje o takovéto fixnosti slova však slovník neobsahuje.

1.1.2 Intersegmenty

Další skupinou výrazů jsou takové idiomy, mezi jejichž slova lze vložit další slovo či více slov. Intersegment (vložené slovo) je typickým způsobem rozvinutí jména (substantiva či adjektiva), který specifikuje účinek idiomu a přináší další zpřesnění.⁴ Délkou intersegmentu (počtem slov) se liší také

³ *bejt klasa*, ale *být klasa*

⁴ *díra do světa* → *velká díra do světa*

idiom od idiomu. V některých případech je možné vložit mezi začátek idiomu a jeho dokončení celou vloženou větou.⁵

1.1.3 Doplnění za neurčitá slova

Zvláštní pozornost je třeba věnovat těm výrazům, jež obsahují neurčitá zájmena či příslovce, za která se v běžné řeči doplňuje konkrétní podstatné jméno. Výrazy, které jsou tímto způsobem nahrazovány, jsou především tyto: *něco, někoho, někomu, něčeho, někým, někam, něčemu, někom, něčem, někde, někudy, něčími, něčímu, něčím, něčích, něčího, něčí, někdo, nějakým, nějaký, nějakéj, nějakou, několika, několikátým, několikátým, několik, někdy, nějak*.

Ve většině případů se zmíněná slova nahrazují, ale existují idiomy, ve kterých může být nějaký z výše uvedených neurčitých výrazů plnohodnotnou významovou součástí hesla. Došlo-li by k jeho nahrazení, výraz se stává neidiomatickým.⁶

1.1.4 Prohození slov

Zvláště u slovesných idiomů se může stát, že sloveso není ve stejné pozici, jako je uvedeno ve slovníku. Tato „přirozená“ vlastnost slovesných idiomů souvisí s jejich postavením přísudku ve větě či souvětí.

1.2 Výrazy neslovesné

Hesla jsou ve slovníku (dle jeho svazků) rozdělena na slovesné a neslovesné. Neslovesné výrazy mají za svůj základ především:

- **podstatné jméno** – *padací most, vlčí morálka,*
- **přídavné jméno** – *až na další, ten pravý,*
- **zájmeno** – *čím dál tím, sám a sám,*
- **příslovce** – *kolem dokola, odshora dolů.*

Jednoznačně nejvíce idiomů je založeno na podstatném jméně s vazbou přídavného jména; nevšední kombinace těchto jmen je základem pro „přenesený“ (idiomatický) význam. Neslovesné idiomy mohou být také založené na jiných slovních druzích, než bylo uvedeno, a to i na číslovkách *páté přes deváté* či na předložkách *na požádání*.

Neslovesné výrazy mají fixnější formu než slovesné, je to dáno především vazbou slov mezi sebou, zatímco u slovesných frází je postavení slovesa a zbytku idiomu navzájem téměř vždy libovolné, což je zapříčiněno volným slovosledem českého jazyka, je většina vazeb u neslovesných idiomů podstatné

⁵ zachovat si chladnou hlavu → zachoval si, i když to bylo velice těžké, chladnou hlavu

⁶ nebýt jen tak někdo, něco nového

– přídavné jméno. Vazba mezi podstatným a přídavným jménem je daleko pevnější, proto zde dochází k vložení intersegmentu či přehození slovosledu jen v menším měřítku.

1.3 Výrazy slovesné

Základem slovesných výrazů je vždy sloveso, jenž je v postavení přísudku věty, avšak toto sloveso nemusí nést hlavní význam idiomu, může jít o neplnovýznamová slovesa, která pouze doplňují plný význam a tvoří tak hlavně gramatickou podstatu věty.⁷ Některé slovesné idiomy vznikly pouhým přidáním slovesa k idiomu neslovesnému, tímto způsobem lze demonstrovat i již zmíněnou relativní nezávislost idiomu na slovese.⁸

Schopnost slovesných idiomů měnit svoji strukturu při zachování idiomatického významu je poměrně vysoká. Je tomu tak proto, že jmenná část, která je většinou druhou součástí slovesné fráze, je oddělitelný od slovesa (přičemž je nadále relativně fixní), a tak vzniká prostor pro umístění slovesa mimo bezprostřední pozici vzhledem k této jmenné části.⁹

U některých slovesných frází je uvedena informace o tom, že se může jedno z jejích slov (sloveso) alternovat pouze s posunem významu slovesa, ale základní sémantický význam zůstane zachován.

1.4 Čištění dat

Zadaná slovníková hesla bylo třeba před dalším zpracováním vyčistit, tj. oprostit je od všech informací, které nebudou při dotazování korpusu zapotřebí. Jednalo se především o vyextrahování holého slovníkového záznamu, tedy lexikálního základu hesla.

Při zpracování byly použity UNIXové nástroje pro práci s textem, především **sed**, **grep**, **tr**, pomocí nichž se podařilo převést zdrojové dokumenty do tvaru, kde je každý holý idiom na samostatném řádku připraven pro další zpracování.

⁷ *mít své meze, být v duchu s někým*

⁸ *díra do světa → udělat díru do světa*

⁹ *Postavit Petra do latě nám dalo práci. vs. Petr se do latě postavit nedal.*

Kapitola 2

Korpus

Hledání výskytů hesel se musí provádět na velkém množství textů, aby výsledná data nebyla zkreslená povahou dokumentu. K vyhledávání je použit korpus SYN2000, který obsahuje více než 100 miliónů tokenů¹, což poskytuje dostatečný základ pro hledání idiomatických frází. Používaný korpus je stejně jako SČFI synchronní, a to převážně z let 1990–1999.^[4]

Slovník, který je dřívějšího vydání, obsahuje hesla, která se v korpusu vůbec nevyskytují. To může mít dvě příčiny:

- idiom se používal pouze ve specifických oblastech, které korpus nepostihuje;
- idiom se používal pouze v určitém časovém období, které se nekryje s vytvářením korpusu.

Ale i fakt, že se idiom v korpusu nevyskytuje, přináší informaci o povaze idiomu. Nevyskytující se idiomy, přinášejí s sebou pravděpodobnost, že jejich významu nebude rozumět.

Slova v korpusu jsou morfologicky označována a lematizována, čehož se využívá při výběru slov, kdy se abstrahuje od morfologické změny slova.

¹Token je nejmenší jednotka zachytitelná v korpusu, jde tedy o slovo, interpunkční znaménko, číslo či jiný nealfabetický symbol.

Kapitola 3

Dotazování

Při vytváření dotazu existují dvě cesty jak a co hledat. První možností je hledání významové podstaty idiomu. V tomto případě se hledá jakýkoli idiom, který lze z neutrálního a obecného slovníkového hesla vytvořit výše popsanými transformacemi. Hledání celé široké škály možných forem, ve kterých se lexikální podoba idiomu může nacházet, nutí k vytvoření množiny možných tvarů pro každý idiom. Způsob, kdy se hledá, do jaké míry je idiom fixní či volný (zda lze prohodit pořadí slov, změnit jejich koncovky či vložit intersegment), je dozajista tou nejlepší (nejpřesnější) cestou, jak najít všechny výskyty idiomu a zároveň nezanést do výsledků žádnou chybu. Vytvoření množiny forem pro každý idiom však není z časového hlediska možné, při zpracování 15 000 výrazů nelze množiny tvořit ručně. Ani počítačové zpracování není při tomto zvoleném přístupu vhodné, protože forma idiomů (zachování idiomatičnosti fráze) je závislá na sémantice, a tak jediné možné správné určení může nalézt pouze člověk, i když i ten nebude schopen obsáhnout všechny možné tvary, protože to s sebou nese předpoklad univerzální znalosti jazyka. Nicméně hledání významového základu idiomu lze využít při hledání všech výskytů několika málo frází, kdy je příprava množin forem zvládnutelná.

Druhou cestou, jak hledat idiomatická hesla, je hledat pouze jejich lexikální základ, tj. zapomenout na významy. Přístup si bere za předpoklad, že blízkost slov obsažených ve frázi je dostatečnou známkou toho, že se idiom v prohledávaném vzorku textu vyskytuje. Je jisté, že, obsahuje-li prohledávaný text hledaný idiom, obsahuje zajisté každé slovo, ze kterého je idiom složen. Na druhou stranu ale neplatí, že, vyskytují-li se slova daného idiomu v textu, bude v textu obsažen i idiom (sémanticky vzato). To s sebou přináší úskalí, protože spolu se všemi výskyty idiomů se nalezne také velké množství výskytů jednotlivých slov, která ale nebudou mít idiomatický význam. To je častou příčinou vysokého počtu nalezených výskytů hesel, která obsahují běžná slova¹. Hlavní výhodou, kterou tento způsob přináší, je, že postihne drtivou

¹ *být, dát, vědět* apod.

většinu forem každého idiomu, stačí tedy jeden univerzální dotaz, do jehož slotů jsou pak jen dosazována jednotlivá připravená slova idiomů. Vytvoření jednoho dotazu a následné generování konkrétních dotazů pro korpus, je velice dobře automatizovatelné, a tím vhodné pro počítačové zpracování.

3.1 Lemmatizace

Při vyhledávání se používá nejen slovo, které idiom obsahuje, ale též jeho lemmatizovaná forma, tj. základní tvar (u podstatných jmen, zájmen a číslovek se za základní tvar pokládá první pád jednotného čísla, u přídavných jmen je navíc přidána podmínka prvního stupně, u sloves je lemmatizovaným tvarem infinitiv a u ostatních slovních druhů je většinou lemma totožné).

Hledání s pomocí lemmatizace abstrahuje od jednotlivých tvarů slov a najít idiomů lišících se v pádě, čísle, rodě, osobě nebo času je jen porovnání získaných lemmat s lemmaty v korpusu.

Korpus je značkován pražským pozičním značkovacím systémem, kdežto idiomy jsou značovány morfologickým analyzátozem *ajka*[3], který byl vyvinut v rámci Laboratoře zpracování přirozeného jazyka Fakulty informatiky Masarykovy univerzity. Tento rozdíl se projevuje nejvíce právě v lemmatizaci, která se pro stejná slova liší.²

K některým, zvláště pak nespisovným, slovům³ nepřihadí *ajka* žádné lemma, proto se v dotazu pracuje nejen s lemmatem, ale také se slovem samotným. Hledání lemmatu a zároveň i slova v původním tvaru řeší jak nepřirazení lemmatu, tak i odlišnou lemmatizaci korpusu a idiomů.

Lemmatizace je ve většině případů víceznačná (pro jeden tvar slova existuje více lemmat)⁴

3.2 Dotazovací jazyk

Pomocí dotazovacího jazyka *Corpora query language* (dále CQL) jsou tvořeny jednotlivé dotazy na korpus. CQL je dotazovací jazyk, pomocí kterého se v korpusu vyhledávají slova dle jejich atributů.⁵ Jazyk dovoluje i kombinování podmínek pro vyhledávání slova. Při hledání slov lze využít této možnosti pro nalezení slova pouze určitého slovního druhu, což se hodí zvláště v analytických jazycích, které mají stejné tvary pro více slovních druhů (angličtina), nebo pro zajištění pokrytí výrazu, pokud není u nějakého slova jeho lemma známo (pokud se lemmatizace nezdaří, je možné se pokusit

²V korpusu má slovo *v* i *ve* stejnou lemmatizaci *v*, kdežto *ajka* přiřazuje slovu *ve* stejné lemma *ve*.

³*vožralej, voprátka aj.*

⁴*žluč* → *žluč*, *žlučit*; *už* → *úžit*, *už*

⁵Záleží na korpusu, které atributy jsou u tokenu uvedeny. Většinou se jedná o word, lemma a tag – morfologické informace.

hledat pouze jeho určitý tvar). Při vytváření dotazu je k dispozici používání regulárních výrazů, které umožňují poměrně velkou flexibilitu v parametrech hledání slova.

Každá pozice (slovo) je ohraničena hranatými závorkami, v nichž se nachází specifikace jednoho nebo více atributů.⁶ Pro každou pozici jsou vytvořeny nové podmínky nebo lze stejné „slovo“ použít vícekrát dodáním počtu opakování za konec hranaté závorky slova, které se opakuje.

Pomocí výše uvedených způsobů hledání jednotlivých slov lze poměrně snadno připravit dotaz na idiom. K popsání struktury většiny neslovesných idiomů totiž stačí nezabývat se morfologickými informacemi. Naopak užití pouze popsaných možností vyhledávání je při dotazování na slovesné idiomy, jejichž pořádek slov není zcela pevný, nedostačující, protože postihnout změny pořádku částí idiomů či vkládání intersegmentů by bylo možné vzhledem k proměnné vzdálenosti mezi částmi pochytit jen velice těžko.

Řešením pro jednodušší vyhledávání slovesných výrazů je použití operátorů `meet` a `union`^[2].

3.2.1 Operátor `meet`

Operátor `meet` má za sebou dva povinné parametry, které odkazují každý na jednu pozici (slovo), další dva parametry jsou nepovinné⁷ a určují, v jakém rozsahu pozic se může druhý parametr (slovo) vyskytovat od prvního. Třetí parametr ukazuje na první možnou pozici druhého slova, čtvrtý pak na poslední, přičemž index prvního parametru (slova) je 0. Nastavením záporných hodnot třetího parametru se dosáhne toho, že se bude druhé slovo hledat před prvním. A to je ona chtěná vlastnost, jež umožňuje postihnout flexibilitu vazeb zvláště u slovesných idiomů.

Jako výsledek aplikace operátoru `meet` je první pozice prvního slova v kontextech, v nichž se objevuje spolu s druhým slovem v zadaném okně, proto lze tento výsledek opět použít jako argument pro vnější operátor. To umožňuje používat vnořené používání operátoru, což je využitelné při hledání slov bez ohledu na jejich pořadí v textu.

Příklad Vyhledání idiomu *dělat/udělat z něčeho aféru*

```
MU (meet (meet (meet [word="dělat|udělat"|lemma="dělat|udělat"]
[word="z"|lemma="z"] -2 5) [word="něčeho"|lemma="něco"] -4 8)
[word="aféru"|lemma="aféra"] -6 11)
```

Pro vyhledání idiomu *dělat/udělat z něčeho aféru* se použil dotaz, který vícenásobně používá operátor v různých kontextech. Nejvnějšší `meet` zpracovává, zda existuje výskyt slova *z* v pozici dvě před až pět za slovy *dělat* či *udělat*. Všechny takovéto výskyty prvního slova, které splnily podmínku,

⁶ [lemma="pes"], [lemma="žrát"&tag="Vp.P.....N...."]

⁷ Pokud nejsou uvedené, jsou implicitně nastaveny na jedna – hledají se slova jdoucí bezprostředně za sebou.

že v jejich blízkosti se nachází dané slovo, se pak stávají prvním argumentem další aplikace operátoru.

3.2.2 Operátor union

Operátor `union` sjednocuje výsledky svých dvou parametrů. To je využito v případech, kdy se alternativní vyjádření idiomu zásadním způsobem liší (*svatá zahrada* vs. *zahrada mrtvých*). V tomto případě je třeba hledat samostatně slovní spojení *svatá zahrada* a nalezené výskyty sjednotit s výskyty fráze *zahrada mrtvých*.

Příklad Vyhledání idiomu *svatá zahrada n. zahrada mrtvých*

```
MU (union (meet [word="svatá"|lemma="svatý|svatá|svatat"]
[word="zahrada"|lemma="zahrada"] -2 5) (meet [word="zahrada"|
lemma="zahrada"] [word="mrtvých"|lemma="mrtvý|mrtvý"] -2 5))
```

3.3 Korpusový manager

Pro přístup ke korpusu byl použit korpusový manager Manatee, jehož řádkové rozhraní `manateesrv` bylo vhodným způsobem využitelné spolu s prostředím `bashe` a UNIXovými textovými nástroji. Pro výpočty četností postačuje zvolené textové rozhraní, které jako výsledek na zadaný korpus a dotaz CQL vrátí jeho četnost.

Kapitola 4

Vytvoření dotazu

Při tvorbě dotazů byl zvolen přístup hledání lexikálního základu idiomu bez ohledu na výsledný sémantický důsledek nalezených frází. Hledání bylo oproštěno o zapojení sémantické informace, kterou může určit pouze člověk, který plně ovládá jazyk. Tento způsob byl zvolen především díky své dobré počítačové zpracovatelnosti a tím pádem i časové zvládnutelnosti.

4.1 Vyhledávání slov

Vyhledávání jednotlivých slov je opřeno o předpoklad, že se slova idiomu mohou vyskytovat v různých morfologických variantách. Proto se při hledání slov nepoužívá pouze atribut *word*, ve kterém je uveden právě původní tvar slova, ale spolu s ním také atribut *lemma*, ve kterém je základní podoba hledaného slova (viz výše). Mohlo by se zdát, že bude stačit pouze lemma, které je pro všechny tvary stejné, ale v některých případech je potřeba použít i atribut *word*, protože se lemmatizace nemusí vždy povést (nespisovná slova či některá citoslovce)¹ nebo lemmatizace přiřadí jiné lemma, než je uvedeno v korpusu², a tak se vyhledává jen pouze podle přesného tvaru, což nemusí postihnout všechny formy idiomu, ale je alespoň nějaká šance, že se najde právě slovníkový tvar.

Lemmatizace slov idiomů je ve většině případů nejednoznačná, což sice přináší další výskyty, které ale s hledaným idiomem nemají nic společného, leč tento problém je odvozen od morfologického analyzátoru a nelze se mu vyhnout.

4.2 Obecný dotaz pro slovesné výrazy

Pro slovesné výrazy je v práci zvoleno hledání pomocí operátoru *meet*, a to především kvůli jejich flexibilní struktuře slov.

¹[word="široka"|lemma=""], [word="zlinkovanej"|lemma=""]

²Korpus: *bejt* → *bejt*, *ajka*: *bejt* → *být*

Dotaz se tvoří rekursivně vzhledem k délce hledaného slovního spojení, kdy se každé další slovo přidává jako druhý parametr operátoru `meet` (prvním parametrem je předchozí slovo nebo výsledek předchozí aplikace operátoru. Při přidávání dalšího slova se automaticky zvětšuje i velikost okna, ve kterém se spojení hledá. Při hledání dvou slov je okno nastaveno v rozmezí -2 až 5, to znamená, že se druhé slovo hledá dvě před až pět za prvním slovem. Tím vzniká dostatečný prostor pro různé variace idiomu, které se liší svojí pozicí. Pro další slova se okno zvětšuje vždy dopředu od dvě pozice a dozadu o tři.

Výsledkem je potom obecný dotaz:

```
(meet ... (meet (meet ([ ] [-2 5] [ ] -2-((w-2-n)*2) 5+((w-2-n)*3))
... [ ] -2-((w-2)*2) 5+((w-2)*3)),
```

kde w je počet slov v hledané idiomatice frázi a n je počet zbývajících slov.

V případě nutnosti lze tyto dotazy posléze spojovat pomocí operátoru `union` a získat tak průnik pro dvě hesla, jejichž strukturu nelze zachytit alternací slov či jejich morfologickými změnami.

Pozice pro slova je vyplněna vždy dotazem, jak na atribut *word*, tak i *lemma*, takže v dotazu se doplní za `[ ]` vyhledání slova s vlastnostmi `[word="wordn" | lemma="lemman"]`, kde *lemma_n* je příslušnou lemmatizací slova *word_n*.

4.3 Obecný dotaz pro neslovesné výrazy

V případě neslovesných hesel byl vytvořen dotaz jiným způsobem. Základním předpokladem bylo, že v neslovesných idiomech jsou vazby mezi slovy pevnější než u slovesných výrazů. Pevnost vazeb je dána především tím, že se idiom nedá rozdělit na dvě nezávisle existující části, které by ve větě mohly stát v různých pozicích. Pevnost jmenných vazeb (z nichž je tvořena převážná část neslovesných výrazů) dokládá jednosložkovost neslovesných idiomů. Proto by použití operátoru `meet` nebylo v tomto případě vhodné, není tu totiž přirozená možnost permutace dvou částí idiomů.³

Pro vytvoření dotazu pro neslovesné idiomy, bylo použito pouze vyhledávání jednotlivých slov, avšak s možností nejvýše jednoho tokenu mezi jednotlivými slovy výrazu. Je to proto, aby byly akceptovány také výrazy, jež vzniknou rozvinutím jednoho ze jmen. Příkladem těchto rozvinutí je přívlastek, který zpřesňuje některé ze slov.⁴ Obecný dotaz pro neslovesné idiomy byl vytvořen následovně:

```
[word="word1" | lemma="lemma1"] [ ] {0,1} [word="word2" | lemma="lemma2"] ...
[ ] {0,1} [word="wordn" | lemma="lemman"]
```

Takto vytvořené dotazy postihují jak morfologickou flexibilitu idiomu, tak též možnost rozvinutí některého ze slov ve výrazu.

³U slovesných výrazů bylo vždy jednou částí sloveso, druhou pozičně nezávislou částí nezávisle na první tvořil zbytek výrazu.

⁴za našich mladých let → za našich nerozvážných mladých let, na přilepšenou → na malou přilepšenou

4.4 Jednoslovná hesla

Struktura dotazu napovídá, že při hledání hesel jsou opomenuta hesla jednoslovná. V části Výrazy neslovesné lze nalézt jednoslovná hesla, která odpovídají abstraktním podstatným jménům, která tvoří kvazifrazémy.[1] Ta jsou ve slovníku uvedena především proto, že je jejich výskyt všeobecně vázán na několik málo kolokačních skupin. Nejde tedy v pravém smyslu slova o idiomatické fráze, ale pouze o upozornění na specifický výskyt některých slov abstraktních⁵. Jde především o specifičnost kontextu, ve kterém se tato slova nacházejí. Slovník uvádí některé možné kolokace, ale vzhledem k datům v korpusu jde pouze o výčet těch nejčastějších. Pro objasnění vazeb těchto slov je lepší použít korpusová data, protože jsou bohatší na možné výskyty vazeb abstraktních substantiv. Četnost takových hesel je rovna četnosti oněch abstraktních slov v korpusu, proto nemá smysl se jimi dále zabývat.

Druhou možností, jak vznikají jednoslovná hesla, je odstranění některých fakultativních částí idiomu (ve slovníku jsou tyto části uvedeny v závorkách) při čištění textu. Je-li původní spojení dvouslovné a jedna z částí není povinnou součástí hesla, pak vzniká po odstranění jednoslovné heslo, jehož četnost je opět rovna pouze jednomu jedinému slovu, nemá proto význam ji počítat znova.

4.5 Vypouštění slov

Stejně jako v předchozím případě, kdy došlo k vypuštění volitelných součástí fráze, se tato redukce počtu slov provádí u víceslovných idiomů a jejich hledání je oproštěno od fakultativních podvýrazů hesla. To je možné právě proto, že je idiom ve své nejkratší podobě neobsahuje, na druhou stranu nic nebrání tomu, aby se vypuštěná část v prohledávaném textu vyskytovala. Umožňuje to právě struktura dotazu, která poskytuje poměrně velká okna v nichž se hledají povinné části idiomů. Mezi povinnými částmi je dostatečný prostor, aby se zde mohlo vyskytovat rozvinutí idiomu či intersegment, což dozajista je vypuštěná část hledané fráze.

4.6 Pokrytí

Pravděpodobnost, že se najdou všechny výskyty hledaného idiomu, je velmi vysoká, protože slova v idiomatické frázi mají určitou pevnost vzájemných vazeb, které nemohou přesahovat hranice vět (pokud ano, tak je to pouze výjimečně).

⁵ *elegance, hra, funus*

4.7 Přesnost

Vzhledem k pokrytí hledaných hesel je přesnost nalezení jen odpovídajících hesel nižší. Je to zapříčiněno především zvoleným přístupem hledání lexikálního základu idiomů. Hesla totiž obsahují taková slova, která se mohou v jiném významu vyskytovat poměrně blízko sebe, aniž se jedná o idiom (drtivá většina slov použitých v idiomech se používá v běžné řeči ve svém základním významu). Dalším znepřesňujícím faktorem je to, že pokud idiom obsahuje neurčité slovo⁶, je toto slovo při hledání nahrazeno libovolným slovem, i když neodpovídá požadavkům na konkretizaci těchto neurčitých slov. Ale ani zajištění podmínky, že hledané slovo bude nahraditelné, není postačující. Idiom si totiž velice často omezuje doménu slov, která lze doplnit.⁷ To jsou hranice, které nelze překročit kvůli sémantické podstatě problému.

⁶*někdo, něco, někde* aj.

⁷U idiomu *vypít něco jedním douškem* je třeba za slovo *něco* doplnit něco, co se dá vypít.

Kapitola 5

Četnost výrazů

Četnost jednotlivých idiomů v jazyku je dána tím, že je výraz vnímán jako aktualizace komunikátu, či jde o automatizovaný jazykový postup, který je běžně používán a v řeči se nad ním příjemce nepozastaví. Mezi automatizované idiomy patří např. spojení *volný čas* či *sníst něco na lačno*. Aktualizace je prováděna především neologickými spojeními, kterými se snaží autor přinést nějakou invenci, mimo další patří mezi idiomy aktualizující *sundat někomu klapky z očí* nebo *starý Procházka*.

Při počítání četnosti idiomu je třeba ohlížet se na původ dat, v nichž se hesla vyhledávají. Nejdůležitější rozdělení je podle toho, zda byl použit jako zdroj mluvený, nebo psaný jazyk. V mluveném jazyce je rozdíl mezi používáním aktualizace a automatických spojení vyšší než u psané podoby jazyka. Je to způsobeno především jednoduchostí používání automatických výrazů, které patří k základním jazykovým znalostem mluvčího, kdežto za aktualizací je skryta intelektuální činnost, která pootvívá další významové roviny komunikátu.

Přesto lze říci, že automatizovaná hesla se používají obecně častěji než aktualizující.

Četnosti dvou základních typů idiomů (slovesné a neslovesné) se liší především kvůli různě zvoleným přístupům k vyhledávání. Přesto lze nalézt některé společné rysy četností, které se vyskytují u obojích výrazů. Je to především fakt, že poměrně velká část idiomů se v korpusu nevyskytuje, u slovesných je to 22 %, u neslovesných výrazů pak necelých 10 %. Dalším společnou charakteristikou četností je postupný pokles počtu idiomů vzhledem k jejich četnosti, nejvíce idiomů má tedy četnost 1–9.

Při zkoumání četnosti hesel lze rozdělit celé spektrum zjištěných četností na tři části:

- **výrazy s četností výskytů nad 1 000** – pokud výsledky hledání výrazů dosáhnou těchto hodnot, nelze s přesností říci, je-li tak dobře, či nikoli. Zvláště v případě slovesných frází se do výsledku započítávají neadekvátní „výskyty“, které dotaz postihl. V případě, že je ale heslo

natolik obecné a aplikovatelné v širokém kontextu (a také ve skutečnosti hojně používané), jsou výsledky správné, což ale značí o tom, že výraz má natolik zanedbatelnou odlišující charakteristiku (viz vymezení výběru idiomů), že jeho používání splývá s běžným použitím jazyka, a je otázkou, zda lze takovéto spojení považovat za idiomatickou frázi¹.

- **výrazy s četností výskytů 10–1 000** – o těchto výrazech lze tvrdit, že představují prototyp ideálního idiomu. Jejich četnost napovídá, že se v jazyku nevyskytují běžně, ale ani zvlášť výjimečně. Výrazy jsou běžnému mluvčímu intuitivně srozumitelné bez potřeby obrátit se na slovník či jinou příručku.
- **výrazy s četností výskytů do 10** – skupina idiomů s touto četností je napříč slovesnými i neslovesnými nejpočetnější (více než polovina všech výrazů). Jde o idiomy, které nemají v psaném jazyku vysoké zastoupení².

5.1 Četnost slovesných výrazů

Slovesné výrazy dosáhly vyšší četnosti než výrazy neslovesné. Tuto skutečnost ovlivnilo několik faktorů, ať už zvolený přístup (struktura dotazu), nebo povaha hesel (automatizované výrazy). Na druhou stranu počet idiomů s nulovým výskytem je u slovesných idiomů dvojnásobný.

Přístup k hledání slovesných frází s sebou přinesl dvě výrazná zvýšení četnosti hesel:

- **použití operátoru *meet*** – kvůli možnosti hledat slova obsažená v idiomu pouze v rámci okna bez jiné další omezující podmínky, se do výsledné četnosti počítají i výskyty, které nepřinášejí žádnou idiomatickou informaci. Je tím porušen princip rozdělení hesla na dvě skupiny – sloveso a ostatní slova, která však drží relativně pevně při sobě, a celý idiom se rozpadá na jednotlivé části po slovech. To přináší největší nárůst četnosti právě u těch idiomů, jejichž lexikální základ se skládá z běžných slov (*být, mít, dát*, jejichž samostatný výskyt je v korpusu velmi vysoký, a proto je zde i velká příležitost k tomu, že v jeho blízkosti se bude vyskytovat druhé (další) slovo z hledaného výrazu).
- **doplňování za neurčitá slova** – toto doplnění umožňuje do idiomu doplnit jakékoli slovo, tzn. pokud idiom obsahuje *něco, někde, někam*, pak je místo hledání těchto slov zařazena „divoká karta“, která vyhovuje jakémukoli slovu. Tímto způsobem pak vznikají spojení, která zákonitě nejsou hledaným idiomem, ale nastavit omezení pro tato slova je velice

¹*volný čas* a další

²V mluvené podobě českého jazyka se může rozložení lišit.

těžké, protože se nelze opřít o význam idiomu a ani o předpoklad, že ono jedno slovo nebude nahrazeno víceslovným spojením či větou.

Kombinací obou předcházejících zvýšení četnosti jde četnost některých idiomů až do řádu deseti tisíců (i výše)³. Jde především o hesla, která se kromě toho, že obsahují velmi časté slovo a slovo s možností doplnění, skládají pouze z předložky, na které pak leží rozlišení toho, co idiom je, nebo není. Řádově vyšší četnost několika desítek idiomů deformuje i průměrnou četnost, která je pro slovesné idiomy 1 323 výskytů pro každý idiom.

A však zjištění, že více než 20 % slovesných idiomů nebylo vůbec nalezeno při zvoleném přesahu hledání je známkou toho, že nenalezené výrazy jsou mimo běžné použití a jejich výskyt je v textu ojedinělý.

Nejzajímavější skupinou idiomů je ta, jejíž četnost se pohybuje v jednotkách až stovkách výskytů v korpusu. O těchto idiomech lze tvrdit, že zvolený přístup byl pro ně správný, tj. potenciál dotazu přesahující strukturu idiomu byl použit adekvátně, získaná četnost odpovídá použití v běžné komunikaci.

Průměrná délka slovesného idiomu je 3,59 slov. Pro hledaných 11 916 slovesných výrazů bylo nalezeno celkem 15 762 893 výskytů v korpusu obsahujícím 114 363 813 tokenů. To by znamenalo, že každé sedmé slovo (13,77 % slov) je součástí nějakého slovesného idiomu. Tento výsledek je poměrně vysoký, ale je třeba brát v úvahu, že přibližně 10 miliónů všech výskytů je tvořeno frázemi *být/bejt někdo*, *být/bejt něco* a *být někoho*. Četnost hesel tím pádem klesá na jednu třetinu.

5.2 Četnost neslovesných výrazů

Neslovesné výrazy se oproti výše popsaným slovesným výrazům liší v četnosti především díky zvolenému systému dotazování spojeného s pevnou vazbou mezi lexikálními jednotkami.

Průměrná četnost výskytů neslovesných výrazů je 191. Tato četnost je sice o řád nižší než u slovesných hesel, zato však není pochyby o tom, že nalezený výraz odpovídá hledanému idiomu. Přesnost je v tomto případě vyšší, což s sebou přináší snížení četnosti hesla.

Průměrná délka neslovesného idiomu je 2,3 slova, je to o více než jedno slovo méně než u slovesných. Odpovídá to nejčastější struktuře hesla, kdy je to kombinace podstatného a přídavného jména.

Při celkové četnosti všech 4 250 zpracovávaných neslovesných idiomatických frází 811 833, je pravděpodobně každé 140. slovo slovem některého z idiomů. To odpovídá 0,71 % všech slov.

Mezi nejčastější neslovesné idiomy patří *jeden z*, *ta jeho*, či *samo sebou*, jež dosahují četnosti výskytů v řádu deseti tisíc a řadí se tak k nejčetnějším. Nad

³Idiomy *být/bejt někdo*, *být/bejt něco*, *být někoho* mají četnost 3 389 707, což dosvědčuje kombinaci častého slova (slovesa) a slova poukazující na jiný objekt

desetitisícovou hranici se dostalo pouze sedm neslovesných výrazů, naproti tomu u slovesných idiomů překročilo tuto hranici 95 výrazů.

Celá jedna čtvrtina výrazů se pohybuje v rozmezí 1–9 výskytů v korpusu. Jde o výrazy jako *babské klepy*, *dušičkové počasí*, *v cuku letu*, *padlá dívka*, *starý kozel* aj.

Četnost neslovesných výrazů je ovlivněna také tím, že mohou být součástí většího celku (lexikálně delšího idiomu). Neslovesné spojení může být jak součástí větší neslovesného výrazu (*soudný den* → *do soudného dne*), tak i slovesného (*druhé housle* → *hrát druhé housle*). Některé výskyty se tak započítávají dvakrát, což zvětšuje četnost idiomů v textu.

Kapitola 6

Další vývoj

Vyvinutí heuristik pro zvětšení přesnosti při zachovaném pokrytí idiomů v textu je jeden z dalších cílů, k němuž by mohlo další prohloubení práce mířit. Přidání syntaktické analýzy hesel by mohlo vést k navržení dotazu, který by byl schopen rozlišovat, zda se výraz v jazyce může vyskytovat i v jiných permutacích slov, než je ve slovníku.

Zapojení syntaktické analýzy by mělo za úkol rozdělit především slovesná hesla na strukturu (*sloveso*, *zbytek výrazu*). Rozdělení idiomů podle této struktury by umožnilo eliminovat permutace jmenných částí slovesných idiomů (*zbytek výrazu*), které obvykle zůstávají pevně spojeny. Implementace systému, který by byl schopen toto zajistit, by přinesla zvýšení přesnosti četností u slovesných hesel. Omezení dotazovacího jazyka, který nedovoluje kombinovat hledání pomocí operátorů *meet* a *union* a výrazů tvořených vyhledáváním jednotlivých slov (jako v případě neslovesných frází), je zarážkou, kterou nelze opomenout.[2]

Dalším možným cílem je vyhledávání obecných víceslovných spojení (*multi-word expression*) v korpusu bez ohledu na jejich idiomaticnost. Toto vyhledávání se může hodit pro terminologická spojení či valence slovesa s konkrétním slovem.

V roce 2009 vyjde nové, přepracované vydání Slovníku české frazeologie a idiomatiky a srovnání výskytů hesel z obou slovníků by mohlo přinést zajímavé rozdíly v četnosti některých hesel. Stejně tak vyhledávání hesel v jiných korpusech by mohlo přinést srovnání a lepší celkový pohled na idiomy v jazyce. Velice přínosné by bylo hledání idiomů v korpusu mluvené češtiny, kde se jazyk zásadně liší, a to především v používání spisovného (nespisovného) jazyka a používání aktualizujících slovních spojení. Výpočet četností v mluveném korpusu by mohl potvrdit či vyvrátit idiomaticnost různých projevů jazyka.

Idiomy jsou jazykově závislé. Jazykové rozdíly se projevují jak v lexikální podobě hesla¹, tak i v existenci napříč různými jazyky. Je to dáno kulturním

¹ *divný brouk* – *queer fish* – *komischer Vogel*

základem jazyka a jeho odlišným vývojem, místními zvyklostmi a reáliemi; idiom jako *španělská vesnice* v češtině existuje významově i v jiných jazycích. V němčině je tento stejný významový základ vyjádřen jako *Böhmisches Dorf* (*česká vesnice*), jde o podobná vyjádření, která jsou ale závislá na tom, v jakém jazykovém prostření je idiom používán (podobně i polský výraz *czeski film*). Přesto srovnání četnosti idiomů v různých jazycích by ukázalo schopnost jazyka vytvářet pevná spojení slov bez ohledu na jejich prvotní význam. Přínosným by mohlo být srovnání především blízkých jazyků (čeština a slovenština) nebo jazyku odlišujících se geograficky (německá, rakouská, švýcarská němčina či britská a americká angličtina).

Vytvořené dotazy na idiomy mohou být použity pro vyhledání příkladů idiomů v korpusu. Využití příkladů, které odkazují na svůj původ v korpusu (nejde jen o vykonstruovaná použití), je žádoucí především při tvorbě frazeologických slovníků a to jak klasických publikací, tak i jejich elektronické podoby, kde je možná přímá návaznost na korpus a příklady mohou být interaktivně vyhledány dle potřeby uživatele.

Závěr

Výsledkem práce je vytvoření dotazů, které jsou schopny postihnout co možná nejvíce variant jednotlivých slovesných a neslovesných frází. Posléze byly tyto dotazy aplikovány na korpus a výsledkem jsou četnosti výskytů idiomů. Při tvorbě dotazů byla zohledněna odlišná struktura slovesných a neslovesných frází, která se liší především flexibilitou vazeb mezi částmi idiomu. Frekvence každého idiomatického hesla ukazuje jeho používání v běžném jazyce. Zatímco hesel, která mají nízké výskyty, je více než polovina, počet hesel, které by měli svojí četností zásadní vliv na jazyk, je pouze malé množství.

Přínos spočtených četností je především v poodkrytí idiomatických vlastností jazyka. Hesla, která nemají svůj kompozicionální význam, se v českém jazyce objevují. Z výsledků plyne, že jejich podíl na jazyce bude kolem 15 %, což by znamenalo velmi vysokou idiomatičnost jazyka, ale to vzhledem k nejčetnějším heslům nelze akceptovat, protože jejich opakovaná četnost tvoří podstatný díl výskytů všech idiomů.

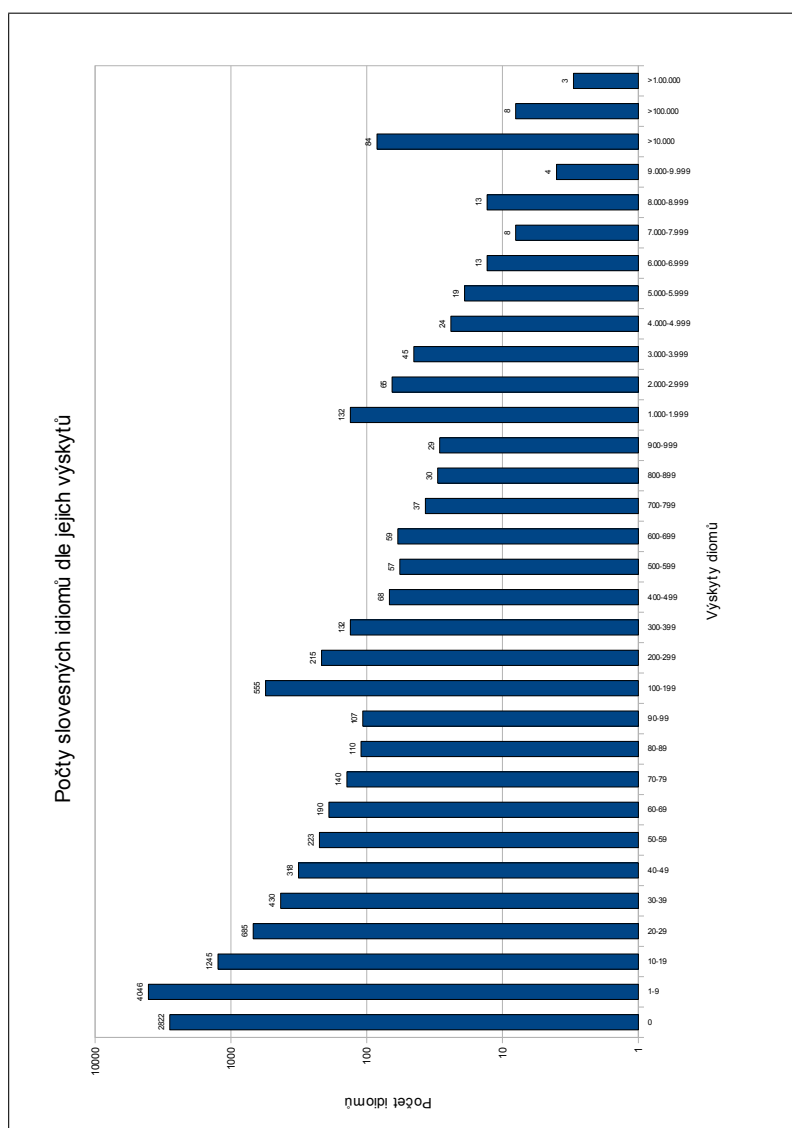
Literatura

- [1] Čermák, František – Hronek, Jiří, editors. *Slovník české frazeologie a idiomatiky*. Praha: Academia, 1994.
- [2] Rychlý, Pavel. *Korpusové manažery a jejich efektivní implementace*. Doktorská práce, Fakulta informatiky, Masarykova univerzita, Brno, 2000.
- [3] Sedláček, Radek. *Morphemic Analyser for Czech*. Doktorská práce, Fakulta informatiky, Masarykova univerzita, Brno, 2005.
- [4] Ústav Českého národního korpusu. *Korpus SYN2000*. [Online; 14. prosince 2008].
- [5] Wikipedia. *Principle of compositionality* — Wikipedia, The Free Encyclopedia [Online; 24. listopadu 2008].

Přílohy

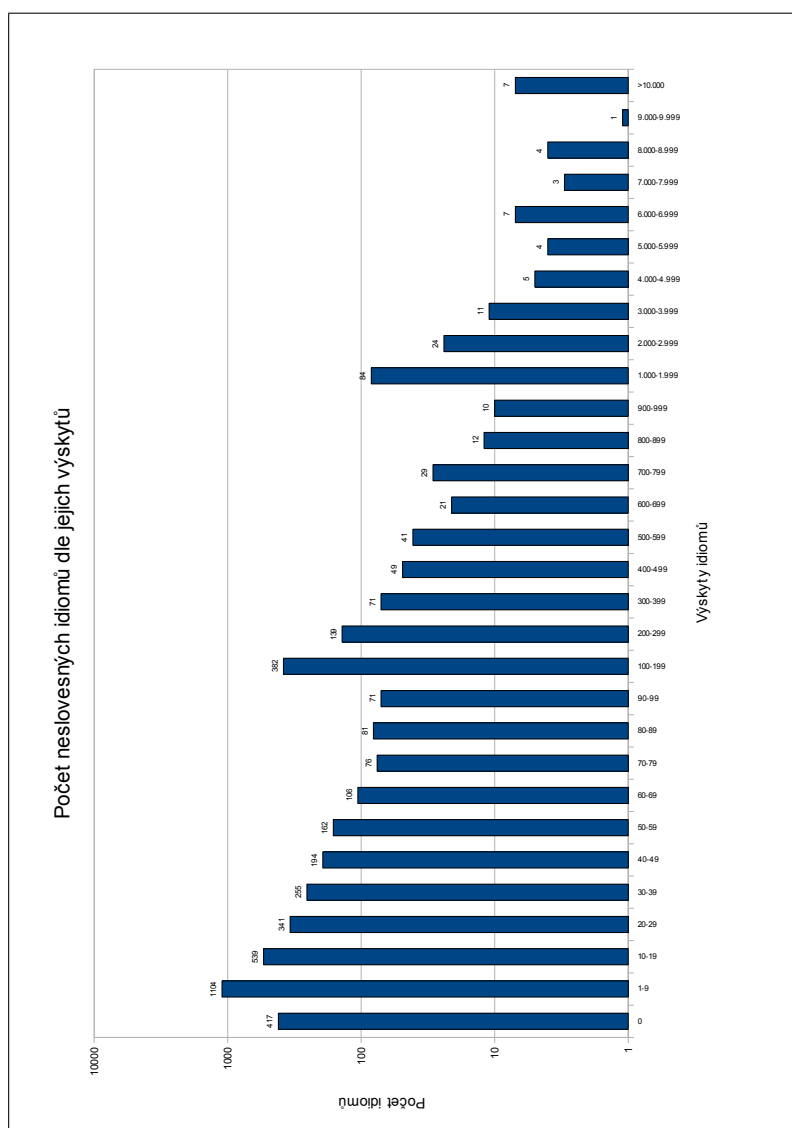
Příloha A

Četnosti slovesných idiomů



Příloha B

Četnosti neslovesných idiomů



Příloha C

Nejčtetnější slovesné idiomy

četnost	idiom	četnost	idiom
3389707	být/bejt někdo	36151	dát/dávat/nechat/nechávat si
3389707	být/bejt něco	35917	nechat/ponechat něco/někoho
3388353	být někoho	35801	mít někoho k něčemu
288037	být/bejt na něčem	29834	mít něco na sobě
288037	být/bejt na něco	28295	vydat/vydávat někoho/něco něčemu/někomu
138499	být/bejt z něčeho	28083	být/bejt při někom
123518	jít/uhnout někomu/něčemu	25851	být celý někdo
107217	být/bejt pro něco	24447	mít někoho za někoho
96502	být/bejt o něčem	24447	mít někoho za něco/někoho
90221	dát/dávat/nechat/nechávat někomu	24447	mít něco za něco
69961	dostat/koupit/slíznout nějakou	23931	být/bejt před někým
69348	mít na někoho	22909	mít z něčeho něco
65099	být/bejt za něčím/někým	22486	být/bejt/stát stranou něčeho
65099	být/bejt za někým	22324	vědět to
65099	být/bejt za něco/někoho	22029	říct něco do/na někoho
64197	být/bejt v tom	21602	ukázat někomu
61636	být/bejt v sobě	19675	být/bejt nic
60098	mít toho/něčeho/někoho	18881	dát/dávat někomu na něco
57285	být/bejt někomu dobrý/dobřej	18578	dát/dávat/nechat/nechávat někomu to
55568	buď jak buď	18006	jet/jít to nějak
50833	mít něco s něčím	17845	být/bejt malým pánem/malý
49175	být/bejt na tom nějak	17044	dělat/udělat si
49146	nebýt na to/něco	16669	jít to nějak
48832	být/bejt po někom	15012	být/bejt proti
48832	být/bejt po něčím	15012	být/bejt proti někomu/něčemu
48832	být/bejt po něčem	14305	jít/chodit na někoho/něco
47542	být/bejt svoji	13542	být/bejt někde znát
47542	být/bejt svůj	13418	provést/vyvést někomu
47528	nebýt svůj	13360	mít v sobě něco
43802	bejt nemožnej	13284	bejt špatnej
41453	být/bejt od něčeho	13156	být něčeho dalek
41273	být/bejt u někoho	12540	přijít na někoho
41273	být/bejt u něčeho	12540	přijít na něco/někoho
40133	dát/dávat si	11977	dát to někomu
36455	dát/dávat/nechat/nechávat se	11977	dát/dávat/nechat/nechávat to/něco

Příloha D

Nejčetnější neslovesné idiomy

četnost	idiom	četnost	idiom
58751	jeden z	3407	veřejné mínění
19215	ten který	3406	v tu chvíli
17271	ta jeho	3383	v žádném případě
15520	samo sebou	3330	na první pohled
15181	na místě	3058	celou dobu
14596	sem tam	2930	cizinecký/cestovní ruch
13612	na světě	2918	západní Evropa
9196	a jiné	2895	před časem
8727	mimo jiné	2894	na čas
8367	v životě	2882	Nový zákon
8256	v podstatě	2802	a podobně
7819	jeden a	2773	samo o sobě
7276	poslední dobou	2773	sám o sobě
7276	poslední dobu	2694	po/na celém světě
6878	ve velkém	2590	na korunu
6789	ze všeho	2536	něco takového
6473	delší čas/dobu	2503	při/po cestě
6367	od minuty/hodiny	2495	v každém případě
6024	v malém	2486	po straně
6022	v poslední době	2385	čtrnáct dní/dnů
5955	trestný čin	2328	v zásadě
5865	delší dobu	2202	z místa
5117	po svém	2066	svým způsobem
5006	v dobrém	2048	dříve/výše/shora uvedený
4764	v duchu	2032	podle mého
4731	valná hromada	2017	všeho druhu
4691	východní Evropa	2009	pracovní síla
4578	něco jiného	1971	ten pravý
4527	ten můj	1954	tento svět
3911	hlavní/rozhodující úloha/role	1949	Bílý dům
3892	na pohled	1934	toho času
3792	v očích	1917	tak trochu
3651	hlavní/rozhodující role	1899	na/v poslední chvíli
3610	celá řada	1867	v minutě
3444	podle svého	1866	více méně

Příloha E

Četnosti všech slovesných a neslovesných idiomů

Na přiloženém CD.

- soubor **frekvence_slovesnych_vyrazu.csv** obsahuje četnost slovesných výrazů oddělenou středníkem od samotného hesla. Záznamy jsou seřazeny podle četnosti hesel.
- soubor **frekvence_neslovesnych_vyrazu.csv** obsahuje četnost neslovesných výrazů oddělenou středníkem od samotného hesla. Záznamy jsou seřazeny podle četnosti hesel.