SÉMANTICKY SIGNIFIKANTNÍ KOLOKACE

Automatická detekce kolokací v českém textovém korpusu

Pavel Pecina a Martin Holub

ÚFAL/CKL Technical Report TR–2002–13

December 2002

Abstrakt

Skupiny slov, která se v textech vyskytují relativně často až typicky společně (v nějakém omezeném kontextu), bývají obecně nazývány *kolokace*. V této práci se soustředíme na tzv. *sémanticky signifikantní kolokace prvního druhu* (dále jen signifikantní kolokace), což jsou slovní spojení vyjadřující specifický význam.

Nejprve rozebíráme definici a charakteristické vlastnosti signifikantních kolokací a navrhujeme jejich klasifikaci z různých hledisek. Dále zkoumáme několik statistických metod automatické detekce dvojslovných signifikantních kolokací v českých textech, srovnáváme je, kombinujeme je, a hodnotíme jejich úspěšnost. Vstupní textová data přitom předzpracováváme s využitím automatické morfologické a syntaktické analýzy textu. Navrhujeme algoritmus pro automatickou konstrukci slovníku signifikantních kolokací. V závěru práce diskutujeme využití signifikantních kolokací jako indexovatelných prvků sémantického slovníku v dokumentografických informačních systémech.

Abstract

Groups of words that occur relatively often or typically together in texts (within a limited context) are commonly called *collocations*. In this work we concentrate on the so called *semantically significant collocations of the first kind* (in the following only significant collocations), which are word phrases that express a specific meaning.

First, we analyse the definition and characteristic features of significant collocations. We suggest several possible classifications of significant collocations from various points of view. Then we investigate several statistically based methods for automatic detection of two-word significant collocations in Czech texts, and we compare, combine and evaluate them. When preprocessing input text data, we make use of automatic procedures for analyzing morphology and syntax. We suggest an algorithm for automatic construction of a dictionary of significant collocations. Finally we discuss possible applications of significant collocations in the framework of information retrieval systems, especially their use as indexable elements of a semantic dictionary.

Obsah

Předmluva

Tato výzkumná zpráva vznikla nepříliš rozsáhlými úpravami diplomové práce [33], kterou vypracoval P. Pecina pod vedením M. Holuba na MFF UK. Původní diplomová práce byla z technických důvodů zkrácena o některé technické dodatky. Originální obsah původní diplomové práce včetně jejích významných datových a softwarových příloh (jež v tomto vydání obsaženy nejsou) lze však najít na webových stránkách Centra komputační lingvistiky (`http://ckl.mff.cuni.cz/ufal/`) v sekci „*Technical reports*“.

Data, použitá pro testování, byla převzata jednak z *Prague Dependency Treebank* verze 1.0 [1], jednak ze zdrojů, které laskavě poskytly redakce časopisu Computerworld a Lidových novin.

Veškerá výzkumná činnost týkající se automatické detekce sémanticky signifikantních kolokací, včetně sepsání a vydání této práce, byla a je finančně podporována projektem MŠMT *„Centrum komputační lingvistiky“* (LN-00A063).

Kapitola 1

Úvod

Kolokace je zajímavý jazykový fenomén. Je to skupina slov, která se v textu vyskytují velmi často nebo až typicky společně. Může to být jakýkoli často používaný a zažitý víceslovný výraz či ustálené slovní spojení. Může to být také skupina významově příbuzných slov, která se velmi často vyskytují ve společném kontextu, říkáme, že *kolokují*.

Kolokace velmi často vytvářejí nový význam, který mnohdy nelze odvodit ze samostatných významů jednotlivých jejich slov - *komponent*, pak hovoříme o *nekompozičnosti*.

Příklady kolokací: *natáhnout bačkory, přijít k sobě, chodit kolem horké kaše, mít navrch, pít krev, udělat rozhodnutí, královská koruna, desetinná čárka, finanční úřad, trestní rejstřík, zbraně hromadného ničení, bílé víno, poštovní novinová služba, nová verze, celé číslo, visutá lanovka, tlustá kniha, dětský lékař, šunka s vejci, vysoký krevní tlak, Karlův most, Sluneční soustava, Josef Novák, letadlo - letiště, auto - silnice, zdravý - nemocný, vysoký - štíhlý . . .*

Z funkčního hlediska jsou kolokace například *idiomatické fráze*, *vlastní jména*, *technické pojmy*, *ustálená slovní spojení* atd. Často jsou na konkrétním přirozeném jazyce značně závislá a jejich překlad do jiného jazyka slovo po slově je obtížný nebo vůbec nemožný.

Přáním odborníků z různých oborů je umět kolokace rozpoznávat. Lexikografové je zařadí do výkladových a překladatelských slovníků. Autoři jazykových učebnic je nezapomenou zmínit při výuce slovíček, jazykoví korektoři při kontrolách pravopisu. Lingvisté je využijí pro rozlišování významů slov či určování větné syntaxe. Pro informatiky mají kolokace význam v oblastech automatického strojového překladu, rozpoznávání textu nebo při jeho syntéze.

Obzvláště důležitým oborem, ve kterém najde rozpoznávání kolokací uplatnění, je vyhledávání informací (*Information Retrieval - IR*) v textech. *Information Retrieval Systems* jsou dokumentografické informační systémy, které na základě uživatelova dotazu, kterým specifikuje, jakou informaci hledá, sestaví odpověď. Touto odpovědí je vybraný seznam dokumentů, které by měly požadovanou informaci obsahovat. Doposud se tento postup nejčastěji prováděl podle shod (měřených různými způsoby) jednoslovných termů v dotazu a ve všech dokumentech. Pokud však budeme moci pracovat místo s jednotlivými slovy také s kolokacemi, můžeme očekávat daleko lepší výsledky. Budeme schopni identifikovat nové významy, které kolokace vytvářejí, přesněji určovat sémantiku uživatelských dotazů a obsah samotných dokumentů.

Cílem této práce je především analýza různých metod automatické detekce různých typů kolokací v českých textech a doporučení metod nejvhodnějších. Nejdříve se budeme zabývat kolokacemi z teoretického hlediska, jejich typickými vlastnostmi a řazením do různých kategorií (v kapitole 2). V kapitole 3 přesně popíšeme řešený problém a nastíníme postup jeho řešení. V kapitole 4 se budeme zabývat předzpracováním vstupních textů, což také zahrnuje šest metod získávání množiny kandidátů na kolokace. Obsahem kapitoly 5 bude popis šesti metod automatické detekce kolokací, které použijeme v našich experimentech. Popis těchto experimentů se nachází v kapitole 6, jejich výsledky pak v kapitole 7. Závěrečné zhodnocení přínosu pro systémy vyhledávání informací je obsahem kapitoly 8.

Kapitola 2

Kolokace

V této kapitole se budeme zabývat kolokacemi z teoretického hlediska. Prodiskutujeme problematiku různých definic tohoto fenoménu, charakteristické znaky kolokací a jejich vlastnosti z různých pohledů. Ukážeme, jak lze kolokace klasifikovat a dle jakých kritérií lze určovat jejich typy. Na závěr se seznámíme s možnostmi uplatnění kolokací v různých odvětvích.

2.1 Vymezení pojmu kolokace

Vymezení pojmu kolokace je poměrně problematickou záležitostí. Kolokacemi, jakožto zajímavým jevem v lingvistice, se v minulosti zabývalo již několik autorů a téměř každý z nich definoval tento pojem jinak. Ani v současné době není definice kolokace ustálená a existuje na ni více názorů. Příčinou je pravděpodobně skutečnost, že samotné určení, zda daná slova tvoří kolokaci, je subjektivní, ovlivněné různými okolnostmi a vůbec nemusí být jednoznačné. Seznamme se tedy s definicemi pojmu kolokace různých autorů.

První, kdo se kolokacemi zabýval byl dle Manninga [6] *Firth*, který v roce 1957 použil tuto definici [2]: „*Kolokace daného slova jsou zažité, tradiční a často používané výrazy, které toto slovo obsahují.*“

Tato definice je poměrně volná, neobsahuje nic o kompozičnosti, syntaktických nebo sémantických vlastnostech. De facto říká, že kolokace se stává kolokací svým častým a zažitým používáním v tradičním významu.

Preciznější definici sestavil *Choueka* v roce 1988 [3]: „*Kolokace je definována jako posloupnost dvou nebo více po sobě jdoucích slov mající znaky syntaktické a sémantické entity, jejíž přesný a jednoznačný význam nebo význam vedlejší nemůže být přímo odvozen z významů nebo vedlejších významů svých komponent.*“. Tato definice je přísnější. Kolokace musí mít syntaktický a sémantický charakter, je již chápána jako lingvistická jednotka (fráze) a poprvé se objevuje znak nekompozičnosti: tzn. význam kolokace nelze zpětně určit z významů jednotlivých slov (viz. kapitola 2.2). Důležitým faktem je Chouekovo omezení, že kolokací musí být *nepřerušená* posloupnost po sobě jdoucích slov.

Významněji se kolokacemi zabývali i *Church* a *Hanks* v roce 1990 [4]. Z Chouekovy definice odstranili omezení nepřerušované sekvence slov. Pracovali však pouze s dvouslovnými koloka-

cemi. Další krok v rozšíření definice provedl *Smadja* v roce 1993 [5]. Odstranil z ní vše, co se týkalo délky kolokací a neomezoval se ani na nepřerušované fráze.

Manning (1999) definuje kolokace podobně široce jako Firth: „ *Kolokace je výraz skládající se ze dvou a více slov, který se používá jako zažité a tradiční označení konkrétního objektu našeho vnímání.*“ Připouští také, že kolokace může být i plně kompoziční. Může se jednat např. o navzájem silně asociovaná slova, která se nemusí vyskytovat v daném pořadí nebo ve společném kontextu. Definice kolokace, která je použita v této práci, vychází z anglického výkladového slovníku *Collins Cobuild English Dictionary* [7]. Pojem *kolokovat (collocate*) dle něj znamená: „vyskytovat se společně“. Kolokace tedy můžeme definovat takto:

> ***Definice 1***: *Kolokace je skupina slov, která se často vyskytují společně.*

Tato definice je však poněkud problematická, zejména použitím dosti nejednoznačných slov *často* a *společně*. Výskyt kolokací nemusí být častý. I její občasné použití nic nemění na tom, že se jedná o kolokaci. Přesněji lze kolokace definovat takto:

> ***Definice 2***: *Kolokace je skupina slov, která se v textu vyskytují relativně často až typicky v nějakém omezeném kontextu společně.*

Ne všechny takto definované kolokace jsou však zajímavé. Např. *být ještě* nebo *moci být* jsou fráze, které jsou velmi používané a tudíž v textech velmi časté. Ze sémantického hlediska jsou ale naprosto nezajímavé. Definujme tedy kolokace sémanticky významné, kterými se budeme dále zabývat.

> ***Definice 3***: *Kolokaci nazýváme sémanticky signifikantní, pokud buď: (A) je to slovní spojení vyjadřující význam, jenž typicky bývá vyjadřován právě tímto způsobem a často je jiným způsobem vyjádřitelný jen obtížně nebo vůbec, anebo (B) obsahuje slova sémanticky související.*
>
> *Slova vytvářející kolokaci nazýváme komponenty kolokace. Sémanticky signifikantní kolokace je tedy struktura konstituovaná svými komponentami a vztahy mezi nimi.*
>
> *Jestliže vztahy mezi komponentami kolokace jsou syntakticko-závislostní (varianta A.), mluvíme o kolokacích prvého druhu – jedná se o idiomatické fráze, odborné termíny mající specifický význam, vlastní jména označující singulární objekty, nebo jiná ustálená a zažitá slovní spojení.*
>
> *V opačném případě (varianta B.), kdy vztah mezi komponentami kolokace je pouze jejich sémantická příbuznost, podobnost, odvozenost nebo jiná sémantická souvislost, hovoříme o kolokacích druhého druhu.*

Poznámka 1: **Kdykoliv bude v celé této práci použit pojem kolokace, budeme mít vždy na mysli kolokace sémanticky signifikantní.**

Poznámka 2: Kolokace prvního druhu v našem pojetí jsou blízké tomu, co bývá dnes často a poněkud vágně označováno jako „fráze“. V tradiční české lingvistice se mluví o tzv. *víceslovném*

nebo *sdruženém pojmenování* nebo o *sousloví*. Citujme např. Šmilauera [26, str. 28]: „... pojmenováním s ustáleným významem a ustálenou formou je *slovo* v širším smyslu slova (i jedno slovo, i sousloví, tj. spojení slov s platností jednoho slova).“; tamtéž, str. 252: „Sousloví (sdružené pojmenování) je ustálené pojmenování, které vzniklo ze spojení dvou i více slov, má však význam slova jednoho a vstupuje do věty jako hotový celek.“.

Problémy s určováním kolokací

Na začátku kapitoly jsme zmínili problematické rozlišování, co kolokace je a co není. Stanovení, zda-li konkrétní n-tice slov je či není kolokací, závisí na mnoha okolnostech:

a) **Doména** - Zatímco technický termín je ve svém oboru jistě kolokací, v jiné oblasti může být zcela kompoziční a kolokaci tvořit nemusí. Kolokace v jedné doméně použití jsou zcela odlišné než v jiné.

b) **Vědomosti** - Úroveň vědomostí a znalostí člověka, který rozhoduje o zařazení sousloví mezi kolokace, tento proces podstatně ovlivňují. Nemusí se jednat pouze o znalosti specifické pro daný obor, ale také o znalosti jazykové.

c) **Subjektivnost** - I kdyby měli dva lidé stejnou úroveň znalostí a vědomostí, vždy se může stát, že jeden z nich sousloví za kolokaci označí a jiný ne. Určování kolokací je vždy subjektivní, což plyne i z povahy samotné definice.

2.2 Charakteristické vlastnosti kolokací

Kolokace jsou velmi zajímavý fenomén a lze u nich pozorovat různé vlastnosti. Podotkněme, že ne všechny kolokace musejí mít všechny následující znaky.

Nekompozičnost

Nekompozičnost je jedním ze základních a nejčastějších znaků kolokací. Pokud je kolokace nekompoziční, znamená to, že její přesný a jednoznačný význam nelze získat pouhým spojením významů jejích jednotlivých komponent. Buď je její význam naprosto odlišný a s významy jednotlivých komponent nemá nic společného (idiomy - *natáhnout bačkory*, *vařit z vody* apod.), nebo se jedná o zvláštní vedlejší či přidaný význam, který není z jednotlivých částí patrný (např. sousloví *bílé víno*, *bílý muž* nebo *bílé vlasy* - ve všech případech je slovem *bílý* míněna zcela jiná barva a nikdy není opravdu bílá. Jedná se o kolokace).

Nesubstituovatelnost

Komponenty kolokací nelze často nahrazovat jinými slovy, i kdyby to byla synonyma nebo svým významem lépe zapadala kontextu. Např. v kolokaci *bílé víno* nemůžeme nahradit přídavné jméno *bílý* za např. *žlutý* i přesto, že může lépe vystihovat barvu bílého vína. Nemůžeme použít frázi *natáhnout přezůvky* místo *natáhnout bačkory* ve významu *zemřít*, nikdo by nám nerozuměl.

Nemodifikovatelnost

Některé kolokace nelze modifikovat tak, že bychom ubrali nebo přidali další lexikální prvek nebo je změnili gramaticky. Toto je opět typický znak idiomů. Tak např. nelze místo fráze *vařit z vody* ve smyslu „mluvit o ničem, vymýšlet si“ použít její modifikovanou verzi *vařit ze studené vody*. Idiom by opět naprosto pozbyl svého původního smyslu. Stejně tak nemůžeme bez újmy na významu říct *zbraně obrovského hromadného ničení* na místo *zbraně hromadného ničení*. Posluchači by nám pravděpodobně rozuměli, ale nač měnit zažitou frázi?

Vnitřní struktura

U kolokací prvního druhu existují mezi jejich komponentami syntaktické závislosti (viz [31], [32]). Jako celek tvoří syntaktický závislostní strom. U dvouslovných kolokací jednoduše rozlišujeme *řídící slovo* (kořen syntaktického stromu) a *závislé slovo*. U delších kolokací je struktura složitější. U kolokací druhého druhu syntaktická závislost mezi komponentami neexistuje, resp. ji neuvažujeme, může být různých typů a je pro nás nepodstatná.

Překlad do jiných jazyků

Častým znakem kolokace bývá skutečnost, že do cizího jazyka se překládá jinak než doslovným překladem každé komponenty. Českou frázi *udělat rozhodnutí* bychom do francouzštiny přeložili slovo po slově jako *faire une décision*, ale správný překlad je *rendre une décision*. Z toho plyne, že *udělat rozhodnutí* je pravděpodobně kolokace.

Dle Šmilauera kolokace mívají synonyma a cizí ekvivalenty jednoslovné (př. *nevím kdo* = *někdo*; *továrna na letadla* = rus. *aviazavod*) [26, str. 252].

Kolokační kontext

Kolokace jsou jevem lokálním. Komponenty kolokace se vždy nacházejí v určité omezené vzdálenosti v textu od sebe. *Kolokační kontext* je kontext slova, ve kterém mohou ležet slova, která s ním tvoří kolokaci. Pro idiomy je tento kontext velmi omezený, často nemohou být modifikovány vložením jiného slova. Ve větě, ve které se vyskytují slova *natáhnout* a *bačkory* jinak než v povrchovém slovosledu těsně za sebou, určitě se nejedná o idiom s významem *zemřít*. Podobně je to u vlastních jmen. Naopak některé kolokace jsou daleko volnější a jejich kontext je větší. Např. kolokace *naklepat maso* se může objevit i ve větě „ *Maminka naklepala včera zakoupené libové vepřové maso.*“, ve které jsou mezi jejím komponentami 4 další slova. U kolokací, které jsou sémanticky podobnými slovy, je kontext kolokace ještě větší, můžeme až překračovat hranice věty, souvislost mezi jednotlivými slovy je pak hůře prokazatelná.

Na kolokační kontext se můžeme dívat z různých pohledů. Může to být větný úsek v povrchovém slovosledu, může to být také část stromu syntaktické závislosti apod. Těchto různých pohledů na kolokační kontext využijeme v kapitole 4.4 jako různé varianty získávání potenciálních kolokací.

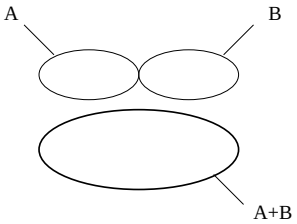
2.3 Klasifikace a kategorizace kolokací

V předchozí kapitole jsme se soustředili na typické vlastnosti kolokací. Nyní si ukážeme, jak lze podle míry projevu těchto vlastností kolokace klasifikovat a kategorizovat. Zařazení nemusí být vždy jednoznačné, některé kategorie se prolínají a hranice mezi nimi není vždy zřetelná.

2.3.1 Sémantické hledisko

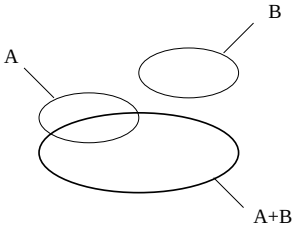
1. *Kolokace dokonale nekompoziční*. Jejich význam je zcela odlišný od významu komponent. Často jsou to idiomy nebo vlastní jména, která neobsahují označení druhu, který pojmenovávají.

   Příklad: *natáhnout bačkory, tlouct špačky, chytat lelky, Oldřich Nový, Coca Cola*

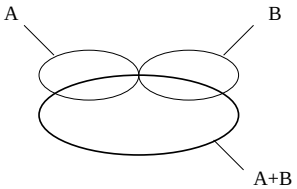


2. *Kolokace částečně nekompoziční*. Alespoň pro jednu komponentu platí, že její význam není zcela obsažen ve významu kolokace (resp. pouze část z významů komponenty je zahrnuta ve významu kolokace). Je to způsobeno především tím, že se jedná o slovo mnohovýznamové (homonymum) a teprve celá kolokace tento význam jednoznačně určuje. Často se také jedná o vlastní jména, která obsahují označení druhu, jenž pojmenovávají (*náměstí, ulice*).

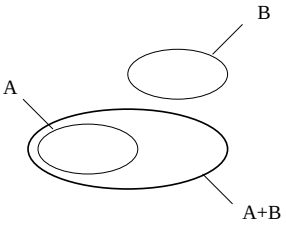
   a) ***Příklad:*** *Národní třída*



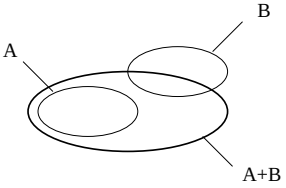
   b) ***Příklad:*** *řádová sestra, Červený kříž*



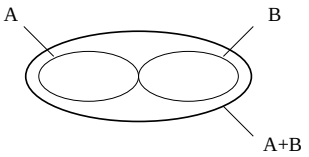
   c) ***Příklad:*** *Karlův most, Staroměstské náměstí*



   d) ***Příklad:*** *koruna stromu, horské oko*

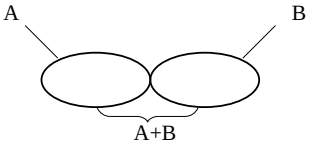


3. *Kolokace slabě nekompoziční*. Význam komponent je zcela obsažen ve významu kolokace, ale nepokrývá jej celý, a stále zde existuje nezanedbatelná „přidaná hodnota“.



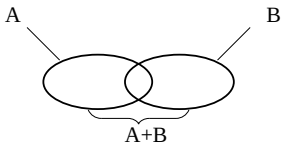
Příklad: *širokoúhlý film, zpětný projektor, řidičský průkaz, odpadkový koš, tisková konference, zubní lékař, znojemská okurka*

4. *Kolokace kompoziční - ustálená slovní spojení bez přidaného významu*. Jejich význam je zcela kompoziční, neobsahuje žádnou „přidanou hodnotu“ jsou však charakteristické častým a zažitým používáním. Komponenty jsou syntakticky závislé.



Příklad: *tlustá kniha, nová verze, nový rok, stručný popis, volné místo, velký objem, minulý rok, tlustá kniha, vysoký stupeň*

5. *Kolokace kompoziční - sémanticky příbuzná nebo blízká slova*. Jejich význam je opět zcela kompoziční, nerozlišujeme u nich žádnou syntaktickou závislost komponent. Jsou to plnovýznamová slova, která se často vyskytují ve stejném kontextu. Tato kategorie je shodná s kolokacemi *druhého druhu*.



Příklad: *kapitán - loď, otázka - odpověď, kapka - déšť, motor - karburátor, židle - stůl, lékař - pacient, černý - bílý, starý - nový*

Ze sémantického hlediska lze kolokace klasifikovat dle vztahu mezi významy dílčích komponent kolokace a významem kolokace samotné. Základní rozdělení je v souladu s definicí kompozičnosti následující: Pokud kolokace vytváří nový význam, tzn. obsahuje jistou „přidanou hodnotu“, kterou nelze získat z dílčích slov, jedná se o **kolokaci nekompoziční**. V případě, že lze význam kolokace získat spojením významu komponent, tzn. žádný nový význam nevytváří, jedná se o **kolokaci kompoziční**.

Pokud bychom chtěli podrobně studovat různé způsoby skládání významu kolokací, můžeme vycházet z výše uvedeného výčtu, který obsahuje tři základní varianty skládání významů nekompozičních kolokací a dvě varianty pro kolokace kompoziční. Uvažujeme kolokace o dvou komponentách, což je pro názornost dostačující; u složitějších kolokací by byl princip stejný. Významy slov i kolokací chápeme jako množiny. Jednotlivé případy odpovídají různým mírám pokrytí či nepokrytí významu komponent významem kolokace. Velký obláček reprezentuje význam kolokace, malý obláček pak zahrnuje význam (nebo významy, pokud se jedná o homonymum) příslušné komponenty.

Jednotlivé kategorie jsou setříděné přibližně podle klesající míry kompozičnosti. Dodejme, že zařazení jednotlivých kolokací do těchto kategorií je subjektivní a může být nejednoznačné, stejně jako pohled na samotnou kompozičnost. Velmi dobrým kritériem pro určení kompozičnosti nebo nekompozičnosti kolokace je následující test.

Test nekompozičnosti kolokace

Nechť někdo zná dobře (dokonale) významy komponent kolokace, ale nezná význam kolokace samotné (třeba ji ještě neslyšel, nebo neumí dobře česky apod.). Pak jestliže mu řekneme kolokaci,

pochopí s jistotou její význam? – Pokud ne, je kolokace (alespoň slabě) nekompoziční. Jinak je (plně) kompoziční.

Rozložitelnost kolokace

Šmilauer [26, str. 252] si všímá toho, zda lze kolokaci rozložit. Citujeme: Sousloví mají různé stupně:

(1) jsou nerozložitelná a jsou v nich slova jinak neužívaná: *křížem krážem, je nabíledni;*

(2) členy jsou těsně spojeny, a proto se přes svou souřadnost neoddělují čárkou: *ve dne v noci, vstávaje lehaje, zuby nehty;*

(3) jsou rozložitelná, ale jako celek metaforická: (motýl) babočka *paví oko* (není ani paví ani oko);

(4) jsou rozložitelná a jen jeden člen je metaforický: *vlčí mák* (je to vskutku mák, ale označení „vlčí“ je obrazné);

(5) oba členy jsou plně významové: *jízdní rychlost*.

2.3.2 Gramatické hledisko

Z pohledu gramatiky lze u kolokací pozorovat dva znaky: z morfologického hlediska **slovní druhy** komponent a z pohledu syntaxe **typy syntaktických závislostí**.

Slovní druhy

Podle slovních druhů jednotlivých komponent u kolokací určujeme tzv. **slovnědruhové typy** kolokací. Jádrem kolokací jsou většinou *autosémantická slova*, tedy slova následujících slovních druhů: N - podstatná jména, A - přídavná jména, C - číslovky, V - slovesa, D - příslovce. V případě delších frází se v nich objevují i předložky a další neautosémantické slovní druhy. Přehled nejčastějších slovnědruhových typů je uveden v tabulce 2.1. Vysvětlivky použitých zkratek jsou také uvedeny v příloze.

Syntaktická závislost

Jak již bylo uvedeno dříve, kolokace mají velmi často syntakticky ucelenou strukturu. V závislostních stromech vět tvoří podstromy, které nemusejí, ale mohou, být dále rozvíjené. Podle druhů syntaktických závislostí můžeme rozlišovat typy dvojslovných kolokací, které jsou uvedeny v tabulce 2.2.

Typ syntaktické závislosti mezi řídícím a závislým slovem můžeme obohatit o slovní druhy obou slov. Každý typ syntaktické závislosti může totiž asociovat slova různých slovních druhů. Získáme tak možnost detailnější kategorizace kolokací. Přehled těchto tzv. *rozšířených typů syntaktické závislosti* je v tabulce 2.3.

typ	příklad
A N	lineární funkce
N N	následník trůnu
D A N	objektově orientovaný jazyk
N A N	zbraně hromadného ničení
N R N	tužka na obočí

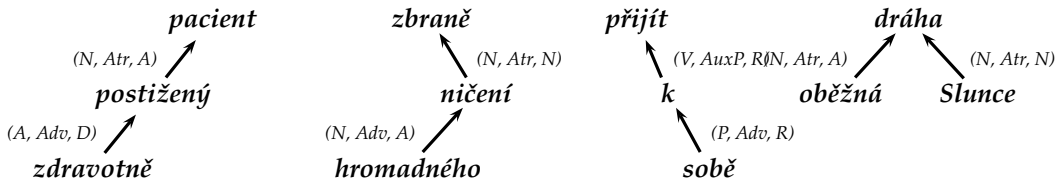
Tabulka 2.1: Nejčastější slovnědruhové typy kolokací

typ závislosti	druh kolokace	příklad kolokace
Atr	přívlastková (atributová)	cenný papír
Sb	podmětová (subjektová)	soud rozhodl
Obj	předmětová (objektová)	dávat přednost
Adv	příslovečná (adverbiální)	zdravotně postižený

Tabulka 2.2: Kategorizace kolokací podle typů syntaktických závislostí

základní typ kolokace	rozšířený typ	příklad
subjektové	(V, Sb, N)	dolar posílil
	(V, Sb, V)	překvapilo, že prší
objektové	(V, Obj, N)	podat hlášení
	(V, Obj, V)	přestalo pršet
	(A, Obj, N)	podobný barvou
atributové	(N, Atr, N)	povrch Země
	(N, Atr, A)	zákony přírodní
adverbiální	(V, Adv, D)	sexuálně obtěžovat
	(V, Adv, N)	poslat poštou
	(A, Adv, D)	silně souvislý
	(A, Adv, N)	poháněný akumulátorem
	(V, Adv, V)	odejít studovat
	(D, Adv, N)	kolmo k přímce

Tabulka 2.3: Typy rozšířených syntaktických závislostí v kolokacích



Tabulka 2.4: Nejčastější typy syntaktických závislostí ve víceslovných kolokacích

U dvojslovných kolokací, kterými se budeme zabývat především, pak můžeme podle druhů závislostí přímo definovat **syntaktické druhy** kolokací. U víceslovných kolokací musíme rozlišovat i tvary jejich závislostních stromů. Příklady jsou na obrázku 2.4.

2.3.3 Pozice v textu

Podle výskytu komponent kolokace v povrchovém slovosledu vůči sobě navzájem rozlišujeme **kolokace pevné** a **kolokace volné**. Pevné kolokace charakterizuje jejich nemodifikovatelnost přidáním dalších slov. Komponenty nelze rozvíjet ani přívlastky nebo příslovečnými určeními tak, že by komponenty kolokace vůči sobě navzájem změnily polohu. Jejich slovosled je neměnný a pevně daný. Volné kolokace toto omezení nemají. Lze je libovolně modifikovat pomocí přívlastků i příslovečných určení, aniž by byl ovlivněn jejich význam nebo aniž by ztratily status kolokace.

Příklad: Pevné kolokace

daňové přiznání, červený kříž, peněžní deník, Velká Británie, přímá viditelnost

Příklad: Volné kolokace

plynová {*elektrická*} *turbína, platební* {*mezibankovní*} *styk, desetinná* {*řádová*} *čárka, spotřeba* {*elektrické*} *energie, klást* {*velký*} *důraz*

2.3.4 Funkční hledisko

Kolokace mohou tvořit různé jazykové útvary (nebo mohou být jejich součástí), např. **idiomy**, **vlastní jména**, **technické výrazy a termíny**. Často obsahují i **významově prázdná slovesa** (*udělat, učinit, být, mít, ap.*), která sama o sobě nemají velký sémantický význam, ale ve spojení s dalšími komponentami kolokace tento význam nabývají (*udělat rozhodnutí*, *mít naspěch*).

2.3.5 Hledisko vyhledávání informací

Z pohledu dokumentografických systémů (DIS) lze dělit kolokace do dvou kategorií. *Kolokace vhodné pro indexování* jsou ty, které má smysl v DIS indexovat, protože tvoří sémantickou jednotku. Jejich výskyty v dokumentech jsou pak evidovány v indexu a je možné je velmi efektivně vyhledávat. Jsou také syntaktickým celkem, který může být součástí uživatelských dotazů. *Kolokace*

nevhodné pro indexování jsou všechny ty, jejichž indexování se z hlediska efektivity vyhledávání v DIS nevyplatí. Takové kolokace nemohou být součástí uživatelských dotazů jako jeden víceslovný term. Dle definice patří do první kategorie všechny kolokace prvního druhu a do druhé kolokace druhého druhu.

2.4 Praktické využití kolokací

Jak již bylo řečeno v úvodu, kolokace mají velmi široké možnosti použití, a to zejména ve dvou oblastech: komputační lingvistice a vyhledávání informací v dokumentografických informačních systémech (DIS). V případě lingvistiky je cílem jednorázově vytvořit co nejobsáhlejší seznam kolokací použitelný v různých oblastech. V případě DIS je cílem kolokace automaticky extrahovat z předložených textů.

2.4.1 Komputační lingvistika

Konstrukce slovníků - ať už sémantických (výkladových) nebo i překladových. Kolokace velmi často vytvářejí nový sémantický význam, jejich automatická extrakce velmi zjednoduší konstrukci slovníku pojmů, termínů atd. Podobně je to i s překlady do cizích jazyků, kolokace často nelze přeložit slovo po slově nezávisle na sobě.

Automatické překlady textů v přirozeném jazyce. Toto využití kolokací opět souvisí s jejich specifickým překladem do jiných jazyků. Pokud budeme mít k dispozici překladový slovník obsahující víceslovné fráze, nebudeme muset překládat slova 1:1, ale N:N nebo M:N.

Kontrola pravopisu. Při kontrole pravopisu mohou být porovnávanými elementy nejen slova, ale i jejich kombinace. Odhalíme tak nejen špatný pravopis slov, ale i případy, kdy jsou celá slova nevhodně použita ve větě (*bílé víno* × *žluté víno*).

Rozpoznávání textu, ale i řeči a jazyka. Obecně jde o jakékoliv případy, kdy se jedná o predikci slov v textu v závislosti na předchozích slovech (tzv. jazykový model). Opět můžeme očekávat větší úspěšnost správného určení slova, které nemůžeme např. zcela rozpoznat, v případě, že předchozí slova svědčí o výskytu konkrétní kolokace.

Syntéza textu, resp. řeči. Jedná se opačný postup k výše uvedenému, tzn. výběr vhodných slov a jejich skládání do větších celků (frází, vět). Použitím slovníku kolokací se můžeme snáze vyvarovat použití nevhodných slov na nevhodných místech.

2.4.2 Vyhledávání informací

Sémantická disambiguace - tedy určování konkrétního významu slova v textu. V mnoha případech tento význam závisí na kontextu a znalost jeho kolokací nebo kolokací ostatních slov zjednodušuje, resp. zlepšuje výsledky tohoto procesu. Kolokace často tvoří slova, která mají více významů, a teprve výskyt těchto slov v kolokaci určuje jejich konkrétní význam.

Klasifikace dokumentů podle jejich obsahu. Jelikož kolokace často vytvářejí nový specifický význam, s jejich znalostí budeme schopni konkrétněji určit téma dokumentu.

Hledání klíčových slov dokumentů. Ze stejných důvodů jako v předchozím případě, pokud v klíčových slovech použijeme kolokace a ne pouze jednoslovné termy, můžeme očekávat, že budou lépe reprezentovat odpovídající dokument.

Indexace dokumentů. V této oblasti je použití pravděpodobně nejpřínosnější. V případě, že jsme schopni z dokumentů extrahovat kolokace, můžeme je spolu s jednoslovnými termy použít jako elementy pro indexaci. Pokud rozpoznáme i kolokace ve vyhledávacích dotazech a použijeme je pro vyhledávání, můžeme opět očekávat výsledky lepší než při klasickém indexování a vyhledávání kombinací slov na sobě nezávislých.

Kapitola 3

Automatická detekce kolokací

V závěru předchozí kapitoly jsme se přesvědčili o velmi širokém uplatnění kolokací. My se soustředíme na jejich využití v oblasti dokumentografických systémů (DIS). Existují tři způsoby, jak kolokace získat:

1. Sestavování seznamu kolokací *lidskou silou*, bez jakékoliv podpory - nesystematické, složité, časově náročné, subjektivní
2. Detekce kolokací z textů *lidskou silou* - systematické, časově náročné, subjektivní
3. *Automatická* detekce kolokací z textů - systematické, časově nenáročné, objektivní, s jistou nepřesností
4. *Poloautomatická* detekce kolokací z textů - systematické, časově náročnější, poměrně přesné, zatížené pouze subjektivností

První dva přístupy jsou problematické tím, že se spoléhají na lidskou sílu, zaměříme se na možnost třetí a pokusíme se eliminovat její nedokonalosti a minimalizovat chyby. Možnost čtvrtá je velmi zajímavá a zmíníme se o ní později. Nyní přesně specifikujme naši úlohu, definujme cíle a navrhněme postup, jak těchto cílů dosáhnout.

3.1 Zadání a specifikace úlohy

Přesné zadání úlohy zní následovně:

> ”Navrhněte a pomocí prostředků výpočetní techniky realizujte a srovnejte metody automatického vyhledávání různých druhů kolokací v českých textech. ”

Podívejme se nyní na jednotlivé části toho zadání a blíže je vysvětleme.

Vyhledávání kolokací v textech. Budeme pracovat s předloženými vstupními texty a v nich se budeme snažit kolokace detekovat. Může se jednat o jakékoliv dokumenty formátované jako

čistý text (*plain text*). Soustředíme se na jejich obsah, nikoli formu. Zpracovávaný text můžeme ekvivalentně označovat jako: texty/dokumenty, kolekce textů/dokumentů, textový korpus apod.

Automatické vyhledávání. Detekce kolokací musí být automatická. Lidskou silou vybrané kolokace můžeme použít např. pro evaluaci a srovnání úspěšnosti různých automatických metod.

Kolokace v českých textech. Dalším předpokladem je, že zpracovávané dokumenty jsou v českém jazyce. V následující kapitole uvidíme, že tato podmínka nemusí být dodržena, a je proto nutné cizojazyčné texty eliminovat. Přísně vzato, naše metody nevyhledávají *české kolokace* (kolokace v češtině), nýbrž *kolokace v českých textech* (česky psaných). Tato distinkce je významná proto, že v českých textech se mohou vyskytnout cizojazyčné výrazy, které jsou ve statistickém zpracování rovnocenné českým, a nelze je automaticky rozlišit. Proto předpokládáme, že ve vstupních zpracovávaných dokumentech k takovému smíšení nedochází, a v následujících kapitolách budeme slovy *české kolokace* rozumět *kolokace v českých textech*, pokud to explicitně nebude řečeno jinak.

Různé druhy kolokací. Kolokace, jak je uvedeno v kapitole 2.3, lze třídit do různých kategorií. Nás zajímá především jejich sémantický význam a proto rozeznáváme kolokace *prvního druhu* a *kolokace druhého druhu*, případně jejich poddruhy dle různých způsobů skládání významu. Využijeme také toho, že různé metody detekce jsou zaměřené na různé druhy kolokací.

Návrh metod. Získávání kolokací není novým problémem (viz kapitola 2.3). Základní metody, jak kolokace detekovat, jsou již známy. Protože se jimi zabývali především anglicky mluvící autoři, jsou přizpůsobeny anglickým textům. Tyto metody se pokusíme upravit pro češtinu, vylepšit, případně navrhnout nové.

Srovnání metod. Abychom mohli různé metody srovnávat, je nutné definovat způsob evaluace, tj. měření jejich úspěšnosti, a stanovit, jaké nástroje k tomu lze použít.

Realizace metod. Ke srovnání úspěšnosti použitých metod je samozřejmě nutná i jejich implementace. Metody implementujeme a pokusíme se je aplikovat na rozsáhlejší textová data a vytvořit tak slovník kolokací.

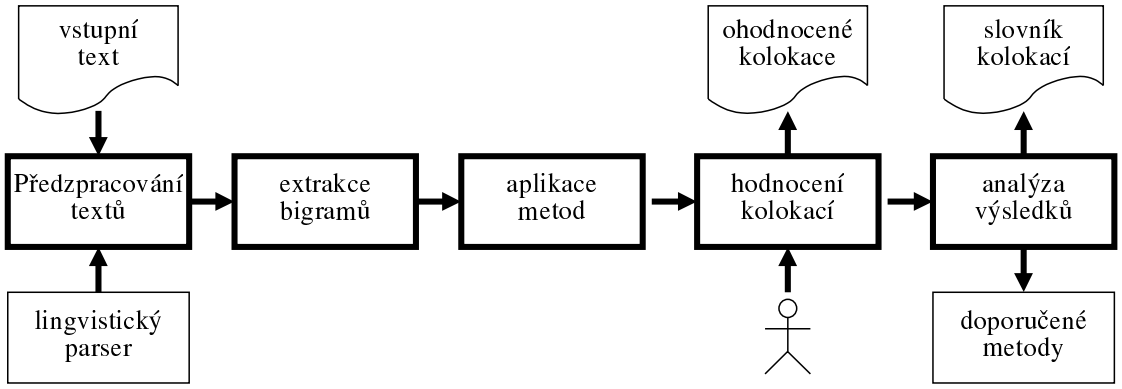
Prostředky výpočetní techniky. Vybrat vhodné prostředky výpočetní techniky je také důležitou stránkou problému, zvláště pokud chceme počítat s nasazením detekce kolokací např. v DIS. Tyto prostředky se navíc mohou lišit pro fázi testování, srovnávání a konečného nasazení.

3.2 Metodologie

Struktura pracovního postupu řešení úlohy je znázorněna na obrázku 3.1. Postup lze rozdělit do několika na sebe navazujících fází:

1. Předzpracování vstupních textů

Předzpracování vstupních textů spočívá v kompletní analýze vstupních dat a odstraňování nejrůznějších chyb, které mohou obsahovat. Data jsou pomocí lingvistické analýzy za použití



Obrázek 3.1: Postup práce při řešení problému detekce kolokací

automatického parseru opatřována podrobnými informacemi o lexikálních jednotkách, morfologických charakteristikách a syntaktických závislostech.

2. Extrakce kolokujících bigramů

Získání kolokujících bigramů jakožto případných kolokací lze realizovat různými postupy. O jaké postupy se jedná, jak je lze vylepšit a jak mohou ovlivnit výsledky - to vše je předmětem fáze č. 2.

3. Aplikace metod detekce kolokací

V tomto kroku provedeme na množinách potenciálních kolokací posloupnost výpočtů a testů. Pro každý bigram pak získáme sadu hodnot, které by měly více či méně vystihnout jeho míru kolokativnosti, tzn. *vlastnosti býti kolokací*.

4. Hodnocení kolokací

Pro srovnání výsledků jednotlivých metod je vhodné mít slovník nezávisle klasifikovaných kolokací - nejlépe lidskou silou. Jak takový slovník získat, případně jak k tomu využít předchozí metody, bude cílem fáze č. 4 – hodnocení kolokací.

5. Analýza výsledků

Poslední fáze spočívá v kompletní analýze výsledků, potvrzení či vyvrácení původních domněnek o specializaci té které metody na ten který druh kolokací apod. Pokusíme se přesněji určit závislosti výsledků jednotlivých metod na tom, zda zkoumaný bigram tvoří či netvoří kolokaci, případně nalézt způsob, jak lze výsledky těchto metod kombinovat a dobrat se tak co nejlepších výsledků. Výsledkem bude aplikace nejlepší metody a sestavení slovníku automaticky detekovaných kolokací.

Kapitola 4

Příprava textů a extrakce kolokujících bigramů

Úvodní fáze procesu automatické detekce kolokací se zabývá předzpracováním a přípravou kolekce vstupních textů. Aby bylo v pozdějších fázích dosaženo co nejlepších výsledků, musí zpracovávané texty splňovat jistá kriteria. Je nutné podrobit je analýze a vybrat pouze ty dokumenty, které daným podmínkám vyhovují.

Český jazyk je z morfologického hlediska velmi bohatý. Například substantiva, která velmi často tvoří součásti kolokací, se mohou vyskytovat až ve čtrnácti různých tvarech (viz tabulka 4.1). Abychom se při detekci kolokací vyhnuli několikanásobnému a zbytečnému zpracování různých slovních tvarů téže slovníkové jednotky, je nutné tyto základní tvary rozpoznat a dál pracovat jen s nimi. Toto se děje na základě výsledků tzv. *lingvistické analýzy*.

4.1 Automatická lingvistická analýza

Lingvistická analýza je součástí širšího procesu, tzv. *značkování*, při němž jsou texty pomocí značek (*tags*) opatřovány doprovodnými informacemi (tj. anotovány)[8]. Text je rozčleněn na hierarchicky uspořádané úseky. Každý úsek je uvozen tzv. otevírací značkou ve tvaru `<značka>`, poté následuje příslušný úsek textu, který bývá ukončen tzv. uzavírací značkou ve tvaru `</značka>` nebo otevírací značkou nového úseku. Každá značka může navíc obsahovat atributy s přídavnými informacemi. Výsledný anotovaný text dodržuje standardy formátu SGML.

Značky jsou trojího druhu: správní (vnější anotace), strukturní a lingvistické (vnitřní anotace). Nejdříve se pomocí *správních značek* vytvoří tzv. hlavička, která obsahuje administrativní údaje o textu. Tvoří ji především informace o původu, autorství, typu a zdroji textu, případně způsobu značkování. Poté již následuje vlastní lingvistická analýza, která probíhá v několika fázích. Schéma celého procesu je znázorněno na obrázku 4.1.

Vlastní text je nejdříve procesem *tokenizace* hierarchicky členěn *strukturními značkami* na menší celky, např. kapitoly, strany, odstavce, ty potom dále na věty, které jsou formálně tvořeny po-

slovní tvar	číslo	pád	c
počítač	S	1	680
počítače	S	2	1 144
počítači	S	3	142
počítač	S	4	465
počítači	S	5	3
počítači	S	6	454
počítačem	S	7	286
počítače	P	1	389
počítačů	P	2	1 742
počítačům	P	3	127
počítače	P	4	664
počítače	P	5	0
počítačích	P	6	422
počítači	P	7	310

Tabulka 4.1: Tvary slova *počítač* a frekvence jejich výskytu v kolekci $\mathcal{CW}94$

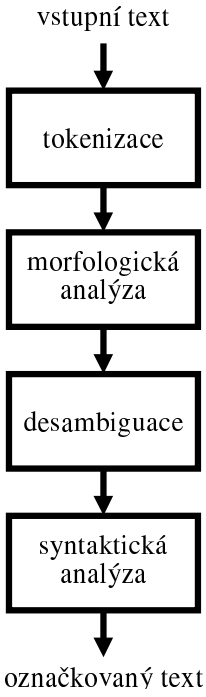
sloupností tzv. textových slov (*tokenů*). Tokeny mohou být slovní tvary, čísla, zkratky, interpunkční znaménka a další zvláštní znaky (symboly měn, fyzikálních jednotek, matematické symboly, atd.).

V fázi *morfologické analýzy* se prostřednictvím *lingvistických značek* přiřazují slovním tvarům jejich možné lingvistické atributy. Konkrétně se jedná o informace dvojího typu:

a) *Lemmatizací* je danému slovu přiřazena informace o jeho slovníkovém tvaru zvaném *lemma*. V případě nejednoznačnosti je přiřazeno více možných základních tvarů (např. *spíš* od slova *spát* nebo *spíše*). Lemma jednoznačně identifikuje slovo jako lexikální jednotku. Je representováno řetězcem písmen a znaků, který odpovídá tzv. slovníkovému tvaru slova, neboli tvaru slova, pod kterým je dané slovo obvykle uváděno ve slovnících.

b) Každé základní formě jsou přiřazeny *všechny* její *potenciální morfologické interpretace*; tj. informace o slovním druhu a dalších morfologických vlastnostech (rodu, čísle a pádu podstatných a přídavných jmen, zájmen a číslovek, o stupni přídavných jmen a příslovcí, o osobě, čísle, slovesném a jmenném rodu slovesných tvarů atd.). Morfologická interpretace je vyjádřena *morfologickou značkou* tvořenou 15 údaji, z nichž každý je reprezentován jedním znakem na příslušné pozici. Význam pozic a znaků na konkrétních pozicích je jednoznačně stanoven. Každá morfologická značka je tedy tvořena patnáctiznakovým řetězcem. Je-li dané slovo morfologicky, lexikálně či slovnědruhově víceznačné, opatří je morfologický analyzátor tolika patnáctiznakovými značkami, kolik má toto slovo lexikálních, slovnědruhových a morfologických významů, a to včetně příslušných lemmat.

Ke každému slovnímu tvaru (dále jen slovu) je tedy přiřazeno jedno či více lemmat a ke každému lemmatu jedna či více morfologických interpretací. Jinými slovy, jedná se o množinu všech teoreticky přípustných interpretací daného slova. Z ní se na základě kontextu vybírá ta nejpravděpodobnější. V konkrétním textu má totiž každé slovo téměř vždy jen jedinou morfolo-



Obrázek 4.1: Schéma procedur automatické anotace textů

gickou, slovnědruhovou a lexikální interpretaci, která nás zajímá především. Tento výběr je cílem procedury zvané *disambiguace* (zjednoznačnění).

Disambiguace je jedním z nejnáročnějších úkolů a největších výzev matematické lingvistiky [8]. Existují dvě základní metody: *stochastická* (statistická, pravděpodobnostní) a *pravidly* řízená.

Stochastická metoda je koncipována na základě stochastického modelu, který je založen na pravděpodobnostech přechodu mezi jednotlivými značkami v morfologicky analyzovaném textu [28], [29]. Tyto pravděpodobnosti se získávají empiricky z ručně označkovaného tzv. *trénovacího* korpusu. Jeho úspěšnost je na úrovni 94 %. Pravidly řízená disambiguace se snaží předchozí způsob vylepšit použitím řady syntaktických pravidel [30].

Poslední fází lingvistické analýzy je proces rozpoznávání syntaktických závislostí ve větách, tzv. *syntaktická analýza* [27]. Jejím výsledkem je závislostní strom, který je v anotovaném textu reprezentován následujícím způsobem: Každý token je jednoznačně označen svým pořadovým číslem ve větě, zároveň je mu přiřazeno pořadové číslo tokenu, na kterém je syntakticky závislý. Pokud token tvoří kořen věty a tudíž není závislý na jiném slově, je mu přiřazena nula. Hrany stromu závislostí jsou ohodnoceny typem příslušné syntaktické závislosti. Každý token pak obsahuje index slova, na kterém je závislý, a zároveň i typ této závislosti.

Na závěr dodejme, že se značkování provádí automaticky pomocí softwarových nástrojů. Podrobnější popis jednotlivých značek a jejich atributů je uveden v přílohách.

Příklad: Označkovaný token

```
<f cap>Nástroje<MDl>nástroj<MDt>NNIP1-----A----<A>ExD<r>1<g>6
```

`<f cap>`	slovo ve tvaru, ve kterém se vyskytovalo v textu
`<MDl>`	lemma s eventuálním upřesněním významu
`<MDt>`	morfologická značka
`<A>`	druh syntaktické závislosti
`<r>`	index tokenu ve větě
`<g>`	index řídícího tokenu

4.2 Chyby vstupních textů

Důležitým předpokladem pro vyhledávání kolokací je dostatek vhodných zpracovávaných textů. Použité metody jsou totiž většinou statistické a kromě zajištění dostatečného objemu dat je nutné brát v úvahu i jejich kvalitu. Ta je ovlivněna různými okolnostmi; jednak chybami, které se mohou vyskytovat na různých úrovních a v různých fázích zpracování, a také samotným obsahem dokumentů. Vyjmutí nevhodných textů zákonitě vede ke snížení statistického šumu, který negativně ovlivňuje celý proces detekce kolokací. Na co se tedy při posuzování kvality dokumentů musíme zaměřit, jaké chyby se v nich mohou objevit a jak se jich můžeme zbavit?

4.2.1 Nízkoúrovňové chyby

Nízkoúrovňovými chybami rozumíme chyby ve vstupních, ještě neanotovaných textech, které znemožní nebo alespoň negativně ovlivní automatickou lexikální analýzu či pozdější fáze vyhledávání. Příkladem tohoto druhu chyb jsou např. nekorektní, nezobrazitelné znaky ve vstupních textových souborech. Projevují se dvěma způsoby: buď dojde k neúspěšnému skončení (pádu) procesu automatického lingvistické analýzy, nebo je tato chyba propagována i do anotovaného textu a podobný pád způsobí až procedurám dalších fází. [1]

Řešení je jednoduché. Ukázalo se, že stačí filtrovat všechny nezobrazitelné znaky a zpracovávat pouze takto vyčištěný text.

Někdy tyto chyby vytváří i samotný lingvistický analyzátor, jehož činnost nemáme možnost ovlivnit. Je proto nutné filtrovat texty před i po lingvistické analýze.

4.2.2 Odlišný jazyk

Dalším problémem, se kterým je možné se při detekci kolokací setkat, je jazyk zpracovávaných textů. Náš úkol je vyhledávání kolokací v českých textech a v průběhu celého procesu počítáme

[1] Ve vstupních souborech se z neznámých důvodů často vyskytoval znak NUL (*binární nula*), který se projevil až tehdy, když byl součástí řetězce v SQL dotazu v databázi, jehož zpracování skončilo chybovým hlášením o neukončeném řetězci. Znak NUL byl totiž při přenosu dotazu do SQL serveru chápán jako jeho konec, i když tomu tak samozřejmě nebylo.

s tím, že pracujeme pouze s českými texty. Pokud by tomu tak nebylo, velmi by se zvýšil v úvodu kapitoly zmiňovaný statistický šum a negativně by ovlivnil výsledky použitých metod.

Nejvíce s předpokladem českých vstupních textů počítá fáze lingvistické analýzy. Morfologický a syntaktický analyzátor jsou určeny pouze pro český jazyk. U ostatních jazyků by úspěšně proběhla snad jen tokenizace, morfologické charakteristiky by byly označeny jako neurčené (nerozpoznané) a výsledky syntaktického analyzátoru jsou v takovém případě nepředvídatelné.

Řešení je tedy dokumenty v jiném než českém jazyce z celé procedury vypustit a zabránit tak zbytečnému zvyšování podílu nerozpoznaných slov.

4.2.3 Odlišná znaková sada

Předpokládejme, že jsme z množiny vstupních textů odstranili všechny, které nejsou napsány v českém jazyce. V oblasti výpočetní techniky však platí, že není český dokument jako český dokument. Problémy mohou vzniknou, pokud je znaková sada těchto textů jiná, než ta, se kterou počítá lingvistický analyzátor (v našem případě *ISO Latin 2*). V souborech s odlišným kódováním se potom opět může vyskytovat příliš mnoho nerozpoznaných slov.

Řešení se nabízejí dvě: buď rozpoznat znakovou sadu každého dokumentu a provést konverzi (ještě před značkováním), nebo, v případě že takto poškozených dokumentů není příliš mnoho, je vyjmout ze zpracování (lze i po procesu značkování).

4.2.4 Nevhodný obsah

I když jsou texty po stránkách jazyka a znakové sady v pořádku, mohou i přesto být nevhodné pro zpracování a detekci kolokací. Problematický může být jejich obsah, druh, zařazení atp. Příkladem mohou být texty, které obsahují tabulky sportovních výsledků,[2] ceny akcií nebo rozsáhlejší citace v cizím jazyce. Opět nejlepším řešením je takové dokumenty nezpracovávat.

4.3 Filtrace vstupních textů

Dokumenty s nevhodným jazykem, znakovou sadou nebo obsahem charakterizuje jedna společná vlastnost. Po lingvistické analýze mají velký nebo alespoň zvýšený podíl nerozpoznaných slovních tvarů (slovní poddruh @ v morfologické značce). Stačí tedy vhodně zvolit maximální podíl těchto slov v dokumentech a ty, které jej překračují, z celého procesu vyjmout.

V úvahu bereme hodnoty dvě: podíl nerozpoznaných slov ve *všech slovech* a podíle nerozpoznaných slov v množině *různých slov*. Ukázalo se, že pokud stanovíme hranici pro obě tyto hodnoty na 15 % a alespoň jedna z nich ji překročí, zbavíme se všech problematických dokumentů a počet neoprávněně vyřazených bude zanedbatelný.

[2] Právě články se sportovními výsledky se často vyskytovaly v kolekci $\mathcal{LN}92$

dokument				nerozpoznaná slova (%)	
jazyk	**kódování**	**kolekce**	**typ**	**všechna**	**různá**
slovenský	ISOLat2	$\mathcal{CW}94$	novinový článek	55,60	61,88
český	Win1250	$\mathcal{CW}94$	novinový článek	35,66	39,55
český	ISOLat2	$\mathcal{LN}92$	zahraniční sportovní výsledky	10,59	19,18
český	ISOLat2	$\mathcal{LN}92$	české sportovní výsledky	7,98	16,14
český	ISOLat2	$\mathcal{LN}92$	novinový článek	1,30	2,23
český	ISOLat2	$\mathcal{CW}94$	rozhovor	1,00	0,75
český	ISOLat2	$\mathcal{PDT}$	kapitola knihy	0,00	0,00

Tabulka 4.2: Příklady podílů nerozpoznaných slov v různých typech dokumentů

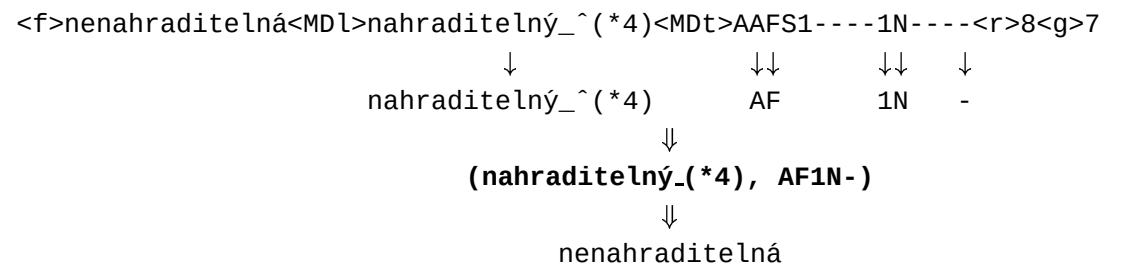
4.4 Metody extrakce kolokujících bigramů

Bigram je obecně jakákoliv dvojice slov. My ovšem budeme jako bigram označovat dvojici slov (tokenů) v základním slovníkovém tvaru. K rozpoznání těchto tvarů využijeme výsledků procesu lemmatizace a morfologické analýzy, které nám pro každý token určují lemma a morfologickou značku, která popisuje jeho konkrétní morfologickou interpretaci. *Základní tvar tokenu* definujeme jako dvojici *(lemma, značka)*, kde *lemma* je lemma tokenu a *značka* vznikla z původní morfologické značky zřetězením hodnot na následujících pozicích:

2. slovní poddruh
3. jmenný rod
10. stupeň
11. negace
15. varianta

Fáze extrakce bigramů se zabývá hledáním tzv. *kolokujících bigramů* jakožto potenciálních kolokací. Způsobů, jak tuto množinu bigramů získat a sestavit, je několik. Kolokací může teoreticky být libovolná dvojice slov a nic nebrání tomu, zvolit tento seznam jako množinu všech dvouprvkových podmnožin množiny všech základních tvarů slov obsažených ve zpracovávaných textech.

Tento způsob získání kolokujících bigramů není však z hlediska dalšího zpracování příliš efektivní. Velikost této množiny by byla $\binom{n}{2}$, kde N je počet různých základních tvarů tokenů. Naším cílem je maximalizovat počet bigramů, které mohou být kolokacemi, a zároveň minimalizovat počet bigramů, které kolokacemi jistě nejsou, a tak se vyvarovat jejich zbytečnému testování. Není tedy nutné uměle vytvářet kombinace slov, která spolu nijak nesouvisí, např. se nikdy nevyskytují ve stejné větě nebo dokumentu. I kdyby v takovém případě kolokaci tvořila, nejsme schopni ji známými metodami rozpoznat. Pro naše výpočty použijeme následujících 6 metod získávání kolokujících bigramů. Vycházejí z různých pohledů na kolokační kontext a různých vlastností kolokací. Pokusíme se je porovnat a doporučit ty nejvhodnější.



Obrázek 4.2: Sestavení základního tvaru tokenu z lemmatu a morfologických značek. Poslední krok je vytvoření snáze čitelného tvaru. Toto zobrazení není však prosté. Z takto vytvořeného slova nelze zpětně získat základní tvar, ale pokud to nepovede k nedorozuměním, budeme je používat pro větší přehlednost.

4.4.1 Slova sousedící

Základní způsob extrakce kolokujících bigramů vychází z teorie, že kolokace je pevná fráze tvořící souvislou posloupnost slov, tzv. *pevná kolokace*. V případě bigramů jsou to dvojice vyskytující se v textu těsně vedle sebe. Vytvořme tedy ze všech textů kolekce posloupnost tokenů, tak jak jdou přirozeně za sebou ve větách, odstavcích a dokumentech (na pořadí dokumentů nezáleží). Každá sousedící dvojice v povrchovém slovosledu pak vytvoří jeden bigram. Počet všech bigramů je $N - 1$, kde N je počet všech slov v kolekci.

4.4.2 Eliminace stopslov

Předchozí metodu můžeme znatelně zlepšit tzv. *eliminací stopslov*. Tato metoda se často používá při *indexaci dokumentů* v DIS. Vytvoří se seznam tzv. *stopslov* obsahující bezvýznamová a málovýznamová slova, která nejsou pro indexaci vhodná. Bývají to především zájmena, předložky, spojky a také některá další velmi frekventovaná slova. Stopslova se poté z procesu indexace vyjmou (viz [9]).

V našem případě použijeme stejný princip. Bigramy obsahující dvě stopslova vyřadíme z dalšího zpracování, neboť kolokace jistě netvoří. Výběr stopslov je ale obtížnější. Nelze do nich zařadit např. některá málovýznamová slovesa jako v případě indexace dokumentů, protože ta ve spojení s dalším slovem kolokace tvořit mohou. Viz následující příklad.

Příklad: Bigramy obsahující málovýznamové sloveso *mít*. Pouze v prvním případě se nejedná o kolokaci, v ostatních zřejmě ano (ale i toto je diskutabilní).

mít *auto*
mít *strach*
mít *náladu*
mít *nedostatek*

Abychom se uvedenému problému vyhnuli, odstraníme všechny bigramy, jejichž obě komponenty nejsou následujících slovních druhů: podstatná jména (N), přídavná jména (A), číslovky (C), slovesa (V) a příslovce (D).

Většina	***práce***	***spočívá***	***v***	***numerických***	***výpočtech***
	Většina práce	*většina spočívá*	*většina v*		
		práce spočívá	*práce v*	*práce numerických*	
			spočívá v	*spočívá numerických*	*spočívá výpočtech*
				v numerických	*v výpočtech*
					numerických výpočtech

Tabulka 4.3: Kolokující bigramy získané posouváním kolokačního okénka velikosti 3

4.4.3 Kolokační okénko

Mějme posloupnost slov $w_n = (w_1, w_2, w_3, \ldots)$. *Vzdálenost slov* w_i a w_j v této posloupnosti je definována jako rozdíl jejich pořadí $i - j$. Tato vzdálenost bývá často označovaná jako *signovaná*.

$$d_s(w_i, w_j) \stackrel{def}{=} i - j \tag{4.1}$$

Pro úplnost definujme také *absolutní vzdálenost slov* w_i a w_j jako absolutní hodnotu jejich vzdálenosti.

$$d(w_i, w_j) \stackrel{def}{=} |i - j| \tag{4.2}$$

Příklad: Vzdálenost slov ve větě

d(analýze, drah) = -3
d(problémů, drah) = -2
d(oběžných, drah) = -1
d(družic, drah) = 1

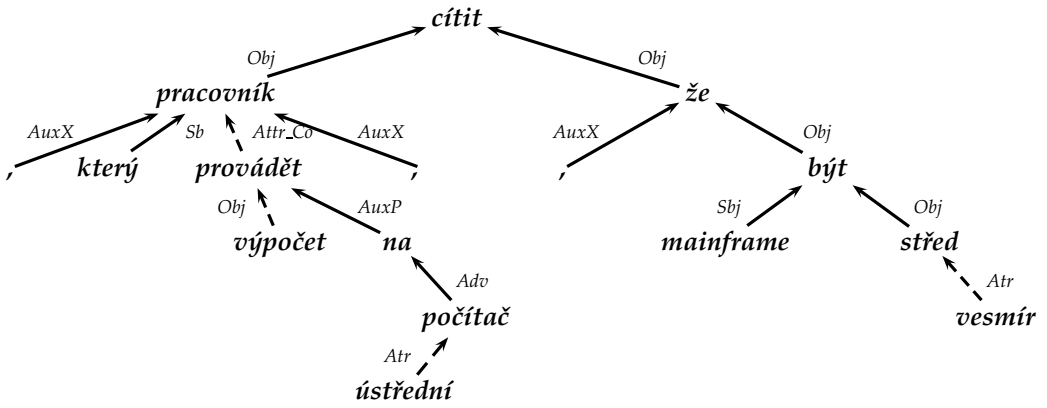
Věta: Nástroj se uplatňuje při analýze problémů oběžných drah družic a v kvantové mechanice.

Předchozí dvě metody jsou vhodné pro detekci pevných kolokací. Existují však kolokace, jejichž výskyt ve větě není pevný a jejich komponenty se tak mohou vyskytovat v různých vzdálenostech od sebe, tzv. *volné kolokace*. Může se jednat např. o kolokaci slovnědruhového typu VN, tedy sloveso a substativum, přičemž substantivum může být rozvito např. různě dlouhým přívlastkem a sloveso příslovečném určením. Tím se vzdálenost komponent stává velmi proměnlivou. Viz příklad:

Příklad: Volné kolokace

a. *Klepat* na *dveře* je zakázané.
b. Jan *zaklepal* na mohutné železné *dveře*.
c. Někdo *zaklepal* znenadání na *dveře*.

Abychom byli schopni detekovat i tento typ kolokací, musíme se soustředit i na širší okolí slov ve větě.



Věta: *Pracovníci, kteří prováděli výpočty na ústředních počítačích, cítili, že mainframe je středem vesmíru.*

Obrázek 4.3: Příklad věty z kolekce $\mathcal{CW}94$ a jejího stromu syntaktické závislosti. Hrany určují dvojice slov, které tvoří bigramy. Ty, jenž lze označit jako kolokace, jsou zvýrazněny přerušovanou čarou.

Definujme pojem *kolokační okénko velikosti* n jako jakýkoliv úsek posloupnosti slov $w_n = (w_1, w_2, w_3, \ldots)$ takový, že pro každou dvojici slov w_i, w_j z tohoto úseku platí $|d(w_i, w_j)| \leq n$.

Množinu kolokujících bigramů získáme tak, že sjednotíme všechny podmnožiny velikosti 2 všech kolokačních okének velikosti n na posloupnosti textů z celé kolekce (n je parametr metody). Jinými slovy: Vytvoříme kolokační okénko na začátku této posloupnosti a všechny dvojice slov z okénka označíme jako kolokující bigramy. Následně kolokační okénko posuneme o jedno slovo dále a opět označíme kolokující bigramy. Proceduru postupně aplikujeme na celou posloupnost. Viz tabulka 4.3.

Takto získaná množina bigramů má však nedostatky. Obsahuje bigramy, jejichž jednotlivé komponenty spolu nijak nesouvisí. Je to v případě, kdy kolokační okénko zasahuje do dvou vět a každá komponenta bigramu se nachází v jiné větě. Výskyty bigramů, které odděluje interpunkční znaménko konce věty, proto vynecháme a snížíme tak i nežádoucí šum.

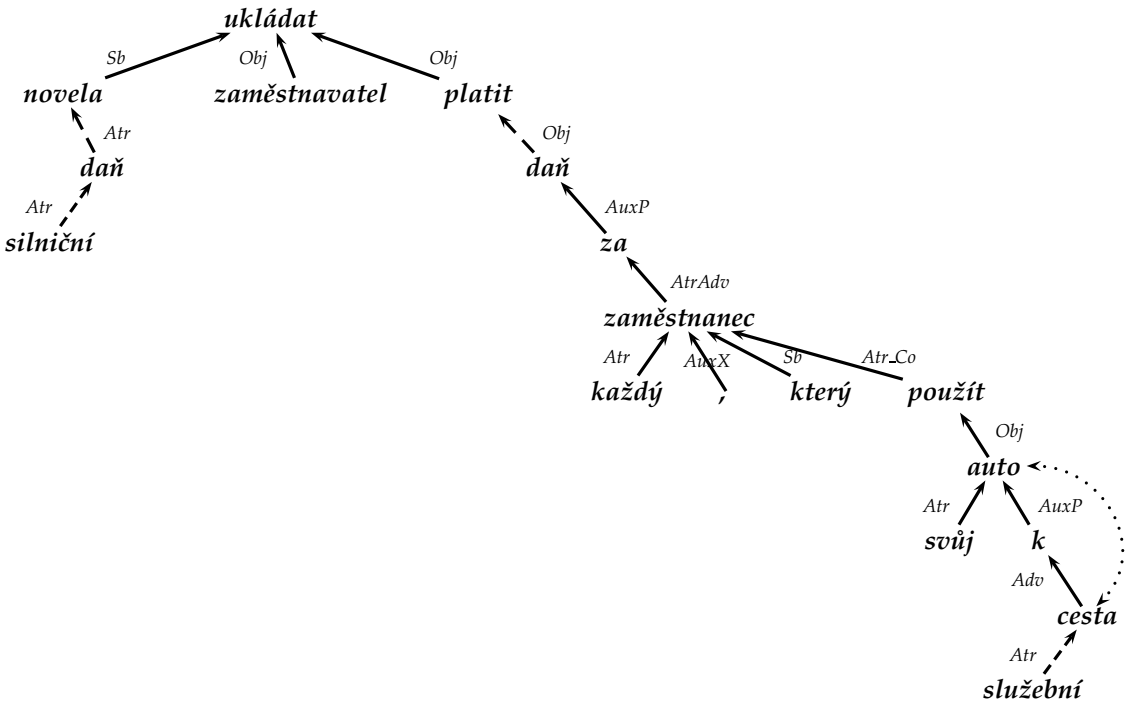
V tomto duchu můžeme zúžit pojem kolokačního okénka a definovat tzv. *kontextové kolokační okénko* jako kolokační okénko, které obsahuje vždy jen slova jedné věty.

Příklad: Bigram zasahující do dvou kolokačních kontextů.

. . . Většina práce spočívá v numerických ***výpočtech****.* ***Opět*** *jsou data v první řadě . . .*

4.4.4 Syntaktická závislost

Nyní se zaměříme na kolokace, které jsou tvořeny slovy, mezi nimiž existuje syntaktická závislost. Sestavme tedy počáteční soubor bigramů jako množinu všech syntakticky závislých dvojic slov, které se v kolekci vyskytují.



Věta: *Novela silniční daně ukládá zaměstnavateli platit daň za každého zaměstnance, který použije své vozidlo k služební cestě.*

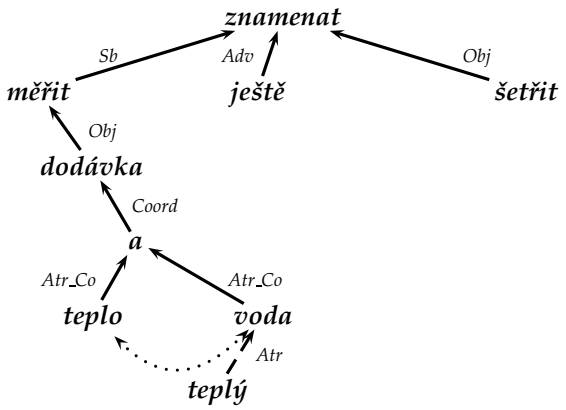
Obrázek 4.4: Stromu syntaktické závislosti. Hrany, které tvoří kolokace syntakticky závislé jsou znázorněny přerušovanou čarou. Tečkovaná čára označuje kompoziční kolokaci s nepřímou syntaktickou závislostí.

Informace o syntaktických závislostech slov ve větě jsou součástí anotovaných textů (viz kapitola 4.1). Pro každou větu můžeme sestrojit tzv. *závislostní strom*, jehož uzly tvoří základní tvary slov, hrany reprezentují syntaktickou závislost a jsou ohodnoceny typem této závislosti, jak je vidět na obrázku 4.3. Relace syntaktické závislosti $\rightarrow$ je asymetrická ve smyslu znázorněném šipkami. Tedy potomek je závislý na svém předkovi - slově řídícím.

Kolokující bigramy odpovídají v tomto stromě jednotlivým hranám.

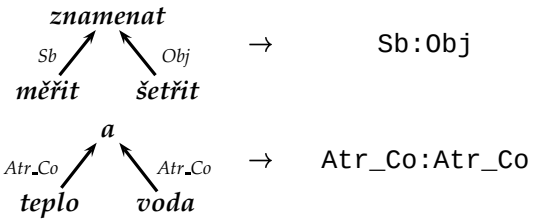
4.4.5 Tranzitivní syntaktická závislost

Na obrázku 4.4 je opět znázorněn závislostní strom věty. Přerušovanou čarou jsou zvýrazněny kolokace, mezi jejichž komponentami existuje přímá syntaktická závislost. Všimněme si také bigramu zvýrazněného tečkovanou čarou *auto – cesta*, který lze označit jako *kompoziční kolokaci*. Tento bigram ovšem zůstane předchozí metodou nezaznamenán. Mezi jeho komponentami totiž existuje pouze nepřímá syntaktická závislost: slovo *cesta* je závislé na předložce *k* a teprve ta závisí na slově *auto*.



Věta: *Měřit dodávky tepla a teplé vody, ještě neznamená šetřit.*

Obrázek 4.5: Strom syntaktické závislosti. Tečkovaná čára označuje kompoziční kolokaci s nepřímou syntaktickou závislostí.



Obrázek 4.6: Vytvoření typu sourozeneckého syntaktického vztahu, konkatenací řetězců, které odpovídají typům syntaktických závislostí obou komponent na společném slově řídícím

Definujme nový typ závislosti, tzv. *tranzitivní syntaktickou závislosti* $\twoheadrightarrow$, tak, že původní relaci syntaktické závislosti rozšíříme o tranzitivitu:

$$x \twoheadrightarrow y \Leftrightarrow \exists z_1, \ldots, z_n \; (x \rightarrow z_1) \wedge (z_1 \rightarrow z_2) \wedge \ldots \wedge (z_{n-1} \rightarrow z_n) \wedge (z_n \rightarrow y) \tag{4.3}$$

Dále pro všechny dvojice slov tranzitivně syntakticky závislých (x, y) definujme tzv. *vzdálenost v závislostním stromu* $d'(x, y)$ jako délku cesty z vrcholu, který reprezentuje slovo x, do vrcholu, který reprezentuje slovo y. [10].

Pomocí předchozích pojmů lze specifikovat novou metodu s parametrem n následovně: Množinu kolokujících bigramů tvoří všechny dvojice tranzitivně syntakticky závislých slov, jejichž vzdálenost v závislostním stromě je menší nebo rovna n.

4.4.6 Sourozenecký syntaktický vztah

Motivací pro poslední metodu extrakce kolokujících bigramů budiž obrázek 4.5. Opět v něm existuje bigram, který můžeme označit jako *kompoziční kolokaci druhého druhu* a který není metodami syntaktické závislosti zaznamenán. Charakterizujícím znakem tohoto a podobných bigramů

je „sourozenecký vztah“ jejich komponent. Formálně: komponenty těchto bigramů jsou syntakticky závislé na stejném slově.

Podobně jako v předchozí metodě definujme s použitím původní relace syntaktické závislosti tzv. *sourozenecký syntaktický vztah* $\swarrow\!\!\searrow$.

$$x \swarrow\!\!\searrow y \Leftrightarrow \exists z(x \rightarrow z) \wedge (y \rightarrow z) \tag{4.4}$$

Počáteční množinu slov sestavíme ze všech dvojic slov, které jsou v této relaci. Součástí každého bigramu bude i *typ sourozeneckého syntaktického vztahu*, který vytvoříme jednoduše konkatenací řetězců, které odpovídají typům syntaktických závislostí obou komponent na společném slově řídícím. Viz schéma 4.6.

Kapitola 5

Metody pro detekci kolokací

V této této kapitole si představíme různé metody automatického vyhledávání kolokací: detekce kolokací podle jejich frekvence v textu, rozpoznávání kolokací založené na střední hodnotě a rozptylu vzdáleností řídícího a kolokujícího slova, testování hypotéz o nezávislosti těchto slov či měření jejich vzájemné informace. Uvedeme si přednosti a také nedostatky těchto principů, které se budeme snažit odstranit nebo alespoň eliminovat. Některé metody rozšíříme a prohloubíme nebo se pomocí heuristik budeme snažit o zlepšení jejich výsledků.

Všechny metody se zaměřují na extrakci dvouslovných kolokací (bigramů), s pevným nebo i volnějším výskytem slov. Základní zpracovávané jednotky textů jsou slova v základním tvaru, podle definice z kapitoly 2. Rozlišujeme slovní druh, mluvnický rod, stupeň, negaci a variantu slovního tvaru. Pro přehlednost výsledků však bude odfiltrována interpunkce, která jako taková nemůže být součástí dvouslovných kolokací (viz kapitola 2). Při detekci kolokací v anglických textech toto zjednodušení nelze použít, protože v případě angličtiny i interpunkční znaménka mohou tvořit kolokace.

Příklady uvedené v této kapitole jsou výsledky jednotlivých metod aplikovaných na kolekci $\mathcal{CW}94$ Jedná se o články z české verze časopisu *Computerworld* z roku 94. Je tedy zřejmé, že i výsledky a příklady budou ovlivněné zaměřením tohoto periodika (informační technologie). Podrobnější informace o této kolekci jsou předmětem přílohy A.

5.1 Frekvence

Nejjednodušší metodou hledání kolokací je počítání jejich výskytů v textu (frekvence). Pokud se totiž dvě slova společně vyskytují častěji, než je obvyklé u jiných dvojic, pak mají pravděpodobně nějaký zvláštní význam, který nelze odvodit z pouhé kombinace těchto slov a jejich společné výskyty se nedají označit za náhodné. Pojem společný výskyt chápeme jako výskyt dvou slov v textu v povrchovém slovosledu těsně za sebou.

Lze však ale předpokládat, že pouhý výběr nejčastěji vedle sebe vyskytujících se slov nebude příliš zajímavý. Viz tabulka 5.1, která ukazuje nejfrekventovanější dvojice sousedících slov v povrchovém slovosledu v kolekci $\mathcal{CW}94$. Vyjma bigramů *operační systém*, *pevný disk*, *pracovní stanice*,

$C(w_1, w_2)$	w_1	w_2
1 595	operační	systém
1 007	být	se
958	být	ten
882	na	trh
880	který	být
796	být	v
740	k	dispozice
734	pevný	disk
731	že	se
690	který	být
689	být	se
664	pracovní	stanice
662	a	ten
646	který	se
644	že	být
634	osobní	počítač
618	a	být
577	se	na
576	a	v
567	informační	systém

Tabulka 5.1: Seznam nejčastějších bigramů v kolekci $\mathcal{CW}94$ a jejich frekvencí. Komponenty bigramů jsou v základním tvaru a pro přehlednost je použito jen jejich lemma. (Bigram *být se* je zde uveden dvakrát. Ve druhém případě se jednalo o sloveso *být* v záporném tvaru.

osobní počítač a *informační systém* se jedná o kombinace funkčních, tedy nikoli významových slov, která pro nás nejsou zajímavá.

Justeson a Katz

Existuje však jednoduchá heuristika, která výsledky této metody velmi výrazně zlepšuje. Jejími autory jsou Justeson a Katz [11] a spočívá v aplikaci tzv. *slovnědruhového filtru*. Tento filtr bere v úvahu slovní druhy jednotlivých komponent kolokací a propustí pouze takové typy bigramů, které vyhovují vzorům vytvořeným podle nejčastějších typů kolokací. Justeson a Katz doporučují vzory uvedené v tabulce 5.2. U každého vzoru je také uveden příklad, který mu vyhovuje. A odpovídá přídavným jménům, P předložkám a N podstatným jménům.

Seznam 20 nejfrekventovanějších bigramů v kolekci $\mathcal{CW}94$, které vyhovují slovnědruhovým vzorům z tabulky 5.2, je uveden v tabulce 5.3. Jak vidíme, výsledky jsou velmi dobré, všechny bigramy lze označit jako kolokace. Protože se zabýváme dvouslovnými kolokacemi, jsou pro nás relevantní pouze první dva vzory.

vzor	**anglický příklad**	**český příklad**
A N	linear function	programovací jazyk
N N	regression coefficients	přenos dat
A A N	Gaussian random variable	síťový operační systém
A N N	cumulative distribution function	softwarový ovladač tiskárny
N A N	mean squared function	kapacita operační paměti
N N N	class probability function	řízení báze dat
N P N	degrees of freedom	tužka na obočí

Tabulka 5.2: Slovnědruhové vzory pro filtrování anglických bigramů použité Justensonem a Katzem

$C(w_1, w_2)$	w_1	w_2	vzor
1 595	operační	systém	AN
734	pevný	disk	AN
664	pracovní	stanice	AN
634	osobní	počítač	AN
567	informační	systém	AN
438	počítač	PC	NN
376	počítačová	síť	AN
333	elektronická	pošta	AN
328	současná	doba	AN
322	síť	LAN	NN
301	nová	verze	AN
294	lokální	síť	AN
291	firma	IBM	NN
281	výpočetní	technika	AN
255	jazyk	C	NN
248	uživatelské	rozhraní	AN
238	datový	typ	AN
231	firma	Apple	NN
225	databázový	systém	AN
208	firma	Microsoft	NN

Tabulka 5.3: Seznam 20 nejčastějších bigramů v kolekci $\mathcal{CW}94$ po aplikaci Justensonova a Katzova filtru

Justensonova a Katzova metoda je velmi přínosná a ukazuje, že i jednoduchá kvantitativní metoda (v našem případě frekvence výskytů) zkombinovaná s nepříliš hlubokou lingvistickou znalostí (důležitost slovních druhů) může být poměrně úspěšná.

$O(w_1, w_2)$	w_1	w_2	vzor
1 666	operační	systém	A N
765	pevný	disk	A N
669	pracovní	stanice	A N
657	osobní	počítač	A N
652	počítač	PC	N N
603	informační	systém	A N
519	systém	firma	N N
505	MS	DOS	N N
388	počítačová	síť	A N
388	počítač	systém	N N
387	lokální	síť	A N
384	síť	LAN	N N
375	firma	IBM	N N
373	systém	DOS	N N
358	novinka	svět	N N
348	produkt	firma	N N
346	počítač	firma	N N
342	nová	verze	A N
337	elektronická	pošta	A N
331	jazyk	C	N N

Tabulka 5.4: Seznam 20 nejfrekventovanějších bigramů se společným výskytem v kolokačním okénku velikosti 5 v kolekci $\mathcal{CW}94$ po aplikaci Justensonova a Katzova filtru. Sloupec označený $O(w_1, w_2)$ obsahuje četnosti dvojic slov ve společném kolokačním okénku.

Kolokační okénko

Výše uvedený postup lze dobře použít pro hledání pevných frázových kolokací. Kolokace, jejichž umístění v textu, resp. větách je volnější, odhaleny nebudou.

S tímto problémem jsme se setkali již v kapitole 4.4 a vyřešili jsme jej s pomocí tzv. *kolokačního okénka*. Nyní zvolíme obdobný postup. Doposud jsme společný výskyt dvou slov chápali tak, že se tato slova v textu vyskytují těsně za sebou, jinými slovy, že jejich absolutní vzdálenost je rovna jedné. Nyní budeme jako společný výskyt dvou slov evidovat všechny případy, kdy absolutní vzdálenost těchto slov je menší nebo rovna n. Jinak řečeno, že se tato slova vyskytují v jednom kolokačním okénku velikosti n.

V tabulce 5.4 je seznam dvaceti nejfrekventovanějších bigramů získaných touto metodou (parametr n jsme zvolili roven 5) po aplikaci slovnědruhového filtru. Nově se v ní objevily následující zajímavé dvojice s příklady jejich konkrétních výskytů:

- ***počítač systém**: počítač se systémem, počítač s operačním systémem*
- ***produkt firma**: produkt zahraniční firmy, produkt světově proslulé firmy*
- ***systém DOS**: systém MS DOS, systém DR DOS*

Použitím *kolokačního okénka* sice bereme v úvahu i slova, která nemají pevné postavení ve větě, ale velká část takových dvojic spolu nemá nic společného, vyskytují se ve své blízkosti zcela náhodně a v těchto výskytech lze těžko hledat příznaky kolokací. V kolekci $\mathcal{CW}94$ se vyskytovala často například tyto skupiny slov: *jak - být*, *teplo - Ostrava* , *již - rok* ale také *telefon - Praha* apod.

Frekvenci bigramů lze využít ještě jedním způsobem. Jak uvidíme v dalších kapitolách, problémy nám budou způsobovat bigramy, které se v textu vyskytují velmi řídce $(1, 2, 3 \dots)$. Můžeme tedy omezit detekci kolokací na bigramy s danou minimální frekvencí výskytu.

5.2 Střední hodnota a rozptyl

Další metodou rozpoznávání kolokací je zkoumání náhodné veličiny vzdálenosti dvou slov v textu, resp. její střední hodnoty a rozptylu, pro jejichž odhady budeme používat známých vztahů z matematické statistiky. Vzdálenosti slov se měří opět pomocí kolokačního okénka, tzn. omezí se maximální možná vzdálenost slov, kterou budeme brát v potaz. Pokud bychom toto omezení nezavedli a uvažovali všechny možné dvojice slov v kolekci, zvýšily by se neúměrně operační nároky na zpracování a nepřineslo by to žádné zlepšení výsledků, spíše naopak. Jak je uvedeno již v kapitole 4.4, kolokace jsou jevem lokálním; je nežádoucí brát v úvahu dvojice slov, které spolu vůbec nesouvisí a jsou od sebe natolik vzdálené, že se nachází v jiných dokumentech nebo větách. Důležité je vhodně zvolit velikost kolokačního okénka n. Manning [6] uvádí jako dostačující hodnotu $n = 4$, pracuje však s anglickými texty. My budeme počítat s velikostí $n = 5$ a ukážeme, jaká hodnota je pro velikost kolokačního okénka pro detekci českých kolokací nejvhodnější.

Vzdálenost slov budeme počítat dle definice 4.1, tedy jako signovanou, nikoli absolutní.

Příklad: Vzdálenosti slova *vytočit* od slova *číslo* ve všech jejich společných výskytech ve stejném kolokačním okénku v kolekci $\mathcal{CW}94$.

a. *Vytočme* tedy spolu *číslo* a požádejme o linku.
b. Počítač *vytočil číslo*.
c. Podesáté *vytočil* stejné telefonní *číslo* ...
d. *Vytočte* před tímto číslem nejdříve *číslo* 001.
e. Dříve, než stačila *vytočit* jeho staré telefonní *číslo*, ...
f. *Vytočil* omylem *číslo* hluché tety.

Odhad střední hodnoty vzdálenosti $\overline{d}$ získáme jednoduše jako aritmetický průměr:

$$\overline{d} = \frac{1}{n} \sum_{i=1}^{n} d_i, \tag{5.1}$$

kde n je četnost společných výskytů, d_i je vzdálenost slov v *i-tém* výskytu.

Příklad: Odhad střední hodnoty vzdálenosti slova *číslo* od řídícího slova *vytočit*.

$$\overline{d} = \frac{1}{6}(3 + 1 + 3 + 5 + 4 + 2) = 3,0$$

Empirický rozptyl s^2 udává, na kolik se vzdálenosti v jednotlivých společných výskytech liší od odhadu střední hodnoty, získáme jej následovně:

$$s^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}. \tag{5.2}$$

kde n opět počet výskytů, d_i vzdálenost slov v i-*tém* výskytu a $\overline{d}$ je střední hodnota, resp. aritmetický průměr. Pokud je vzdálenost ve všech případech stejná, potom je rozptyl roven nule. Pokud je náhodná (což je znakem toho, že společný výskyt je náhodný), pak je rozptyl velký.

Příklad: Odhad rozptylu vzdálenosti slova *číslo* od řídícího slova *vytočit*.

$$s^2 = \frac{1}{5}((3-3)^2 + (1-3)^2 + (3-3)^2 + (5-3)^2 + (4-3)^2 + (2-3)^2) = 2,0$$

Další veličinou, kterou budeme používat, je směrodatná odchylka s. Určuje proměnlivost vzdálenosti zkoumaných slov. Je rovna druhé odmocnině z s^2.

$$s = \sqrt{s^2}. \tag{5.3}$$

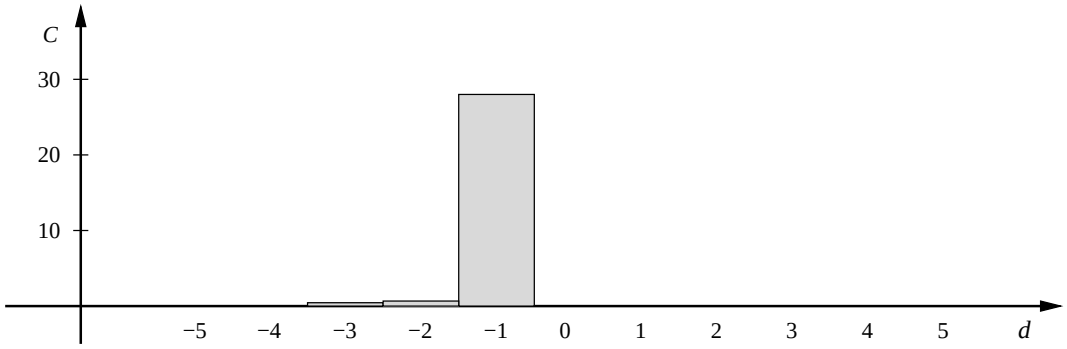
Příklad: Směrodatná odchylka vzdálenosti slova *číslo* od řídícího slova *vytočit*.

$$s = \sqrt{2} \approx 1,414$$

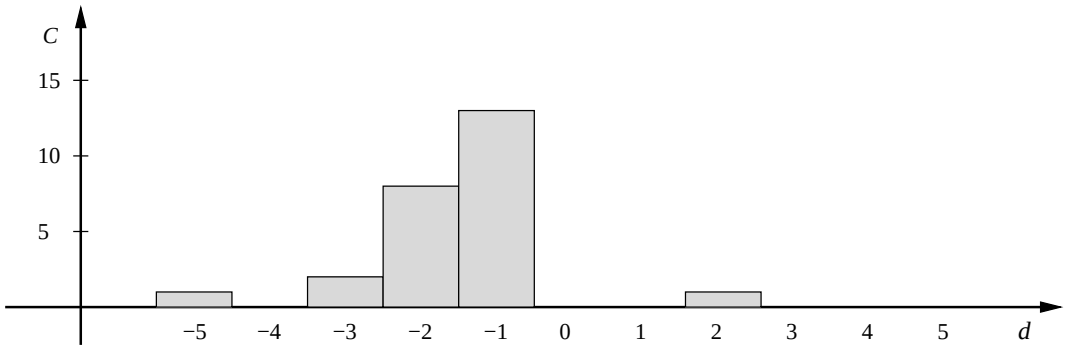
Nyní si ukážeme, jak lze střední hodnotu a směrodatnou odchylku vzdálenosti slov v textu využít pro detekci kolokací. Jednou z možností je hledat bigramy s malou směrodatnou odchylkou. Takové dvojice slov se obvykle vyskytují v přibližně stejné vzdálenosti od sebe. Pokud je směrodatná odchylka nulová, znamená to, že se zkoumaná slova vždy vyskytují ve stejné vzdálenosti.

Můžeme také zkoumat rozdělení naměřených vzdáleností dvojic slov, neboli frekvence výskytů pro jednotlivé signované vzdálenosti. Obrázek 5.1 znázorňuje tři případy, které nás zajímají především.

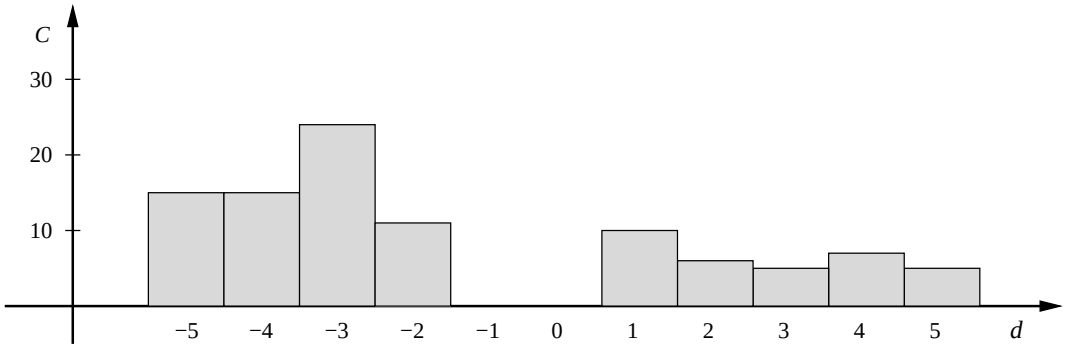
V tabulce 5.5 jsou příklady kolokací a jejich frekvencí, které odpovídají třem právě popsaným variantám. Pokud je aritmetický průměr blízký nule a odchylka také malá, jak je tomu např. u sousloví *pracovní stanice*, pak se jedná o kolokaci, kterou objevila i metoda Justensona a Katze. Pokud je aritmetický průměr podstatně větší než jedna a odchylka malá, pak se pravděpodobně jedná o vhodného kandidáta na kolokaci, kterou ale metoda Justensona a Katze neměla šanci objevit. Kolokace *legenda obrázku* se může vyskytovat ve frázích např. *legenda* {*prvního*} *obrázku*



Pozice slova *odborná* vzhledem ke slovu *veřejnost* ($\overline{d} = -1,12$, $s = 0,42$)



Pozice slova *automatizovaný* vzhledem ke slovu *systém* ($\overline{d} = -1,52$, $s = 1,19$)



Pozice slova *nový* vzhledem ke slovu *systém* ($\overline{d} = 1,40$, $s = 3,21$)

Obrázek 5.1: Histogramy vzdáleností

s	$\overline{d}$	$O(w_1, w_2)$	w_1	w_2
0,42	-0.96	669	pracovní	stanice
0,13	-2.02	58	legenda	obrázku
0,44	-2.12	25	uvést	trh
0,34	-2.89	23	pravá	strana
3,93	-0.21	42	firma	model
4,15	-0.16	62	rok	systém
3,93	-0.80	40	současný	firma
3,73	-0.25	28	jiný	síť
1,01	-1.07	105	laserová	tiskárna
1,08	-1.50	10	poštovní	služba
1,02	-2.04	23	čtecí	hlava
1,13	-2.57	7	typ	závislosti

Tabulka 5.5: Detekce kolokací pomocí odhadu střední hodnoty a rozptylu. Střední hodnota $\overline{d}$, směrodatná odchylka s a frekvence c pro 12 bigramů.

nebo *legenda* {*výše uvedeného*} *obrázku*; *uvést trh* se může vyskytovat ve frázích např. *uvést* {*na*} *trh* nebo *uvést* {*na východo-evropský*} *trh* atd.

Velká směrodatná odchylka značí, že zkoumaná slova se nevyskytují v žádném zajímavém vztahu a kolokace to s největší pravděpodobností nejsou. Potvrzují to příklady čtyř dvojic slov s největšími směrodatnými odchylkami v tabulce 5.5. Pozornost však zasluhují bigramy, které svojí hodnotou směrodatné odchylky spadají mezi tyto dvě kategorie (třetí část tabulky 5.5), tedy se zvýšenou odchylkou a navíc vysokou frekvencí výskytu. Znamená to, že se často vyskytují v několika pevných vzdálenostech od sebe. Opět se jedná o zajímavé kandidáty na kolokace. Např. dvojice slov *poštovní služba* může být částí těchto frází: *poštovní služba*, *poštovní novinová služba* atp.

Metodu detekce kolokací založenou na měření střední hodnoty, kterou jsme se v této kapitole zabývali, popsal poprvé Smadja v roce 1993 [5]. Pokusil se ji také rozšířit tak, že odfiltroval všechny bigramy, jejichž histogramy vzdáleností měly „plochá“ maxima, tedy maxima, která byla obklopena podobnými hodnotami. Příklad takového nevhodného maxima vidíme na obrázku 5.1 u dvojice *nový systém* v bodě -3. Toto vylepšení se ukázalo být velmi úspěšné pro detekci *terminologických frází* (poměr správně určených kolokací byl odhadem asi 80 %) a frází využívaných při generování přirozeného jazyka.

Jak bylo uvedeno již v kapitole 2, Smadjova definice kolokace je nejbližší té naší, je mnohem volnější než definice jiných autorů. Je to dáno právě Smadjovou metodu, která detekuje i kolokace, které ostatní autoři ani za kolokace nepovažují. Např. *získat respekt* je určitě kompoziční, nevytváří nový význam, ale přesto je důležitá např. v oblastech generování textů či překladů. Slovo *získat* má několik synonym (*dostat*, *obdržet*, *nabýt*), ale ve spojení se slovem *respekt* se používá pouze ono. Proto dvojici *získat respekt* mezi kolokace řadíme.

$C(w_1, w_2)$	$w_1 w_2$
107	celý systém
69	počítač typ
68	systém firma
64	nový počítač
50	jiné slovo
49	jiný počítač
47	počítač firma
41	typ počítač
41	nový typ
39	další možnost

Tabulka 5.6: Bigramy s velkým počtem výskytů a malým rozptylem vzdáleností slov, které nejsou kolokacemi. Jejich častý společný výskyt je pouze náhodný, způsobený tím, že jejich komponenty se samostatně vyskytují také velmi často.

Metody založené na měření střední hodnoty a rozptylu jsou tedy vhodné především pro detekci kolokací s méně pevnými vazbami mezi slovy, než jaké jsou u pevných frází, a jejichž umístění v textu není pevně dané.

5.3 Testování hypotéz

Až do této chvíle jsme se vyhýbali jednomu důležitému faktu, a sice že vysoká frekvence bigramu a malý rozptyl vzdálenosti jeho komponent mohou být způsobeny zcela nahodilým a nezávislým výskytem těchto slov vedle sebe. V této a následující kapitole budeme text chápat jako posloupnost náhodně vybraných slov. V případě, že jsou některá slova sama o sobě v této posloupnosti velmi častá, můžeme očekávat i větší počet případů, kdy se budou vyskytovat v této posloupnosti za sebou, aniž by mezi nimi existovala nějaká závislost. To naznačuje, že ne všechny vysoce frekventované dvojice slov jsou pro nás zajímavé. Příklady takových bigramů vidíme v tabulce 5.6: *celý systém* a *nový počítač*.

Potřebovali bychom tedy posoudit, zda společný výskyt dvou slov je zcela nahodilý nebo je zapříčiněn nějakým vztahem mezi těmito slovy. Jedná se vlastně o jeden z klasických problémů statistiky, tj. určení závislosti nebo nezávislosti jevů. Obvykle se tento problém řeší tzv. *testováním hypotéz*, které probíhá v těchto krocích:

1. Nejdříve se formuluje tzv. *nulová hypotéza* H_0 (výrok) o pozorovaných jevech. V našem případě jsou těmito jevy výskyty slov tvořící bigram a hypotézou tvrzení, že neexistuje žádná souvislost mezi těmito výskyty, tzn. že jsou nezávislé [12]. Naším úkolem bude rozhodnout, zda tuto hypotézu můžeme pro jednotlivé bigramy zamítnout či nikoliv. Jako *alternativní hypotézu* H_1 označíme negaci hypotézy H_0.

2. Určí se *hladina spolehlivosti* $1 - \alpha$, kde α je pravděpodobnost zamítnutí H_0, v případě že H_0 platí. Za dostačující se běžně považuje hladina spolehlivosti $0,95$ ($\alpha = 0,05$), avšak

v případě, že pracujeme s velmi velkým objemem dat, což je v NLP běžné, lze hladinu spolehlivosti stanovit na $0,995$, ta odpovídá hodnotě $\alpha = 0,005$.

3. Testujeme hypotézu H_0 pro konkrétní náhodný výběr z rozdělení náhodné veličiny na hladině spolehlivosti $1 - \alpha$ oproti H_1. Toto se děje výpočtem testové statistiky, která je specifická pro každý druh testu.

4. Pro každý test existuje tzv. *kritický obor*, což je množina hodnot výběrů, pro které hypotézu H_0 zamítáme (*obor zamítnutí hypotézy*). Této množině odpovídá tzv. *kritická hodnota*, se kterou srovnáváme testovou statistiku z bodu 2. Pokud je tato statistika větší než příslušná kritická hodnota, hypotézu H_0 na hladině spolehlivosti $1 - \alpha$ zamítáme, v opačném případě ji nezamítáme.

Ukažme si, jak můžeme tento princip použít pro rozhodování, zda dvojice slov tvoří kolokaci či nikoliv. Budeme studovat následující jevy (N je délka zkoumané posloupnosti):

- w: výskyt slova w v textu. Můžeme měřit četnost takových výskytů $C(w)$ v textu. Potom maximální věrohodný odhad pravděpodobnosti, že náhodně vybrané slovo je w, je $P(w) = \frac{C(w)}{N}$.
- $w_1\, w_2$: společný výskyt dvou slov w_1, w_2 v textu. Míníme tím situaci, kdy se slova w_1, w_2 nacházejí na pozicích i a $i + 1$, tedy těsně za sebou. Můžeme měřit počet těchto výskytů $C(w_1 w_2)$ a maximální věrohodný odhad pravděpodobnosti, že náhodně vybrané slovo je w_1 a další náhodně vybrané slovo je w_2, je $P(w_1 w_2) = \frac{C(w_1 w_2)}{N}$.

Zkoumat můžeme i jevy doplňkové:

- $\neg w$: výskyt všech ostatních slov v textu kromě slova w. Četnost $C(\neg w) = N - C(w)$ a maximální věrohodný odhad pravděpodobnosti, že náhodně vybrané slovo není w, je $P(\neg w) = \frac{C(\neg w)}{N}$. Zároveň :

$$P(\neg w) = \frac{C(\neg w)}{N} = \frac{N - C(w)}{N} = 1 - \frac{C(w)}{N} = 1 - P(w)$$

- $\neg w_1\, w_2$: společný výskyt slov uv, kde $u \neq w_1$ a $v = w_2$
- $w_1\, \neg w_2$: společný výskyt slov uv, kde $u = w_1$ a $v \neq w_2$
- $\neg w_1\, \neg w_2$: společný výskyt slov uv, kde $u \neq w_1$ a $v \neq w_2$

Nulová hypotéza H_0 popisuje situaci, kdy zkoumaný bigram netvoří kolokaci. Tzn., že společný výskyt jeho komponent je nahodilý, neboli, že výskyt slova w_1 je nezávislý na výskytu slova w_2 a opačně, tedy že odhad pravděpodobnost společného výskytu slov w_1 a w_2 je součinem odhadů pravděpodobnosti výskytu slova w_1 a pravděpodobnosti výskytu slova w_2. Formálně:

$$H_0 : P(w_1 w_2) = P(w_1)P(w_2) \tag{5.4}$$

Tento model je zjednodušený. Ve skutečnosti nelze říci, že text v přirozeném jazyce je posloupností náhodně vybraných slov. K tomuto problému se ještě vrátíme v závěru kapitoly.

Testování hypotéz se provádí statistickými testy. Těch existuje několik a pro naše účely se hodí zejména *test t*, χ^2 *test* a *poměr pravděpodobností*.

t	$C(w_1)$	$C(w_2)$	$C(w_1, w_2)$	w_1	w_2
4,4720	25	25	20	červený	kříž
4,4719	20	98	20	řádová	čárka
4,4719	50	31	20	San	Francisko
4,4691	32	686	20	aritmetická	operace
4,4682	77	364	20	podavač	papíru
1,3980	9 781	2 274	20	systém	typu
1,3969	21 394	1 040	20	na	další
0,9743	5 300	4 775	20	program	v
0,9730	769	32 922	20	technika	být
0,7272	823	32 922	20	úroveň	být

Tabulka 5.7: Výsledky aplikace *t testu* na bigramy z kolekci $\mathcal{CW}94$ s frekvencí právě 20

5.3.1 t test

t test je pro detekci kolokací velmi často používaný. Při jeho aplikaci se sleduje střední hodnota a rozptyl vybraného vzorku měření, přičemž nulová hypotéza zní, že rozdělení tohoto výběru má určitou střední hodnotu μ, která odpovídá situaci, kdy jsou výskyty nezávislé. Test zkoumá rozdíl mezi očekávanou a empiricky naměřenou střední hodnotou váženou empirickým rozptylem dat a říká nám, na kolik je možné, že vzorek s touto empirickou střední hodnotou a rozptylem je náhodným výběrem z normálního rozdělení, střední hodnotou μ. Používá se k tomu následující testová statistika:

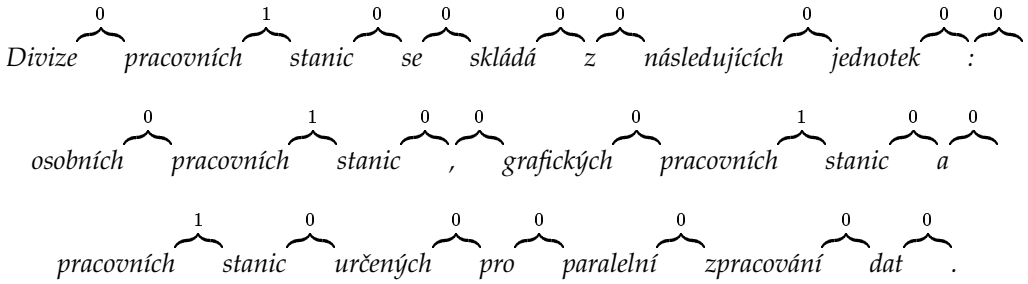
$$T = \frac{|\overline{x} - \mu|}{\sqrt{\frac{s^2}{N}}}, \tag{5.5}$$

kde $\overline{x}$ je empirická střední hodnota zkoumaného vzorku, s^2 je empirický rozptyl toho vzorku, N je velikost vzorku a μ je předpokládaná střední hodnota rozdělení. Pokud je hodnota T dostatečně velká nulovou hypotézu zamítneme. Přesnou kritickou hodnotu T pro zamítnutí hypotézy lze vyhledat ve statistických tabulkách. Závisí také na tzv. *stupních volnosti*, což je hodnota $N - 1$. V našem případě, kdy zkoumáme velký objem dat, bereme stupně volnosti jako ∞.

Použití *t testu* pro vyhledávání kolokací ukážeme na příkladu výpočtu hodnoty t pro bigram *nová společnost* v kolekci $\mathcal{CW}94$. Použijeme standardní rozšíření *t testu* pro práci s celými čísly.

Zpracovávaný vstupní text, ve kterém hledáme kolokace, budeme chápat jako posloupnost náhodně generovaných slov o velikosti $N + 1$. Na tuto posloupnost se můžeme také dívat jako na posloupnost N náhodně generovaných bigramů. Vzorek pro danou dvojici slov, kterou chceme zkoumat získáme jako posloupnost hodnot náhodné veličiny (označme ji X_t), jejíž hodnoty korespondují se zmíněnou posloupností bigramů. Tato veličina nabývá hodnoty 1, pokud se na odpovídající pozici nachází sledovaný bigram, a 0 v ostatních případech. Příklad konstrukce tohoto vzorku (posloupnosti hodnot náhodné veličiny X) je na obr. 5.2.

Počet všech tokenů v kolekci $\mathcal{CW}94$ je $N = 1\,617\,845$, slovo *nová* se v něm vyskytuje 1 608-krát a slovo *společnost* 9 781-krát. Spočteme maximální věrohodný odhad pravděpodobnosti výskytů



Obrázek 5.2: Příklad sestrojení posloupnosti hodnot náhodné veličiny $X_t(\textit{pracovní stanice})$ pro *t test*

slov *nová* a *společnost* založený na pozorování odpovídajícího vzorku:

$$\begin{aligned} P(\textit{nová}) &= \frac{1480}{1617845} \\ P(\textit{společnost}) &= \frac{2747}{1617845} \end{aligned}$$

Nulová hypotéza H_0 je tvrzení, že výskyty slova *nová* a *společnost* jsou nezávislé.

$$\begin{aligned} H_0 : P(\textit{nová společnost}) &= P(\textit{nová})P(\textit{společnost}) \\ &= \frac{1480}{1617845} \times \frac{2747}{1617845} \\ &\approx 1,55327 \times 10^{-6} \end{aligned}$$

Pokud je nulová hypotéza pravdivá, potom náhodná veličina $X_t(\textit{nová společnost})$, jejíž realizací jsme získali 0-1 posloupnost pro bigram *nová společnost*, má Bernouliho rozdělení s odhadem pravděpodobnosti:

$$\hat{p} = 1,55327 \times 10^{-6}$$

Střední hodnota pro toto rozdělení je:

$$\mu = p \approx 1,55327 \times 10^{-6}$$

a rozptyl:

$$\sigma^2 = p(1-p) \approx p$$

Aproximace $p(1-p) \approx p$ je použita z důvodu, že p je ve většině případů velmi malé a jeho druhou mocninu lze tedy zanedbat.

Počet společných výskytů *nová společnost* je 9 a maximální věrohodný odhad pravděpodobnosti společného výskytu je:

$$p' = \frac{9}{1617845} \approx 5,56296 \times 10^{-6},$$

tedy odpovídající odhad střední hodnoty:

$$\overline{x} \approx 5,56296 \times 10^{-6},$$

$O_{11} = \|\|w_1, w_2\|\|$	$O_{12} = \|\|\neg w_1, w_2\|\|$
$O_{21} = \|\|w_1, \neg w_2\|\|$	$O_{22} = \|\|\neg w_1, \neg w_2\|\|$

Tabulka 5.8: Kontingenční tabulka χ^2 testu při odhalování závislosti mezi výskyty slov w_1 a w_2

	$w_1 = každá$	$w_1 \neq každá$
$w_2 = firma$	$O_{11} = 6$	$O_{12} = 8478$
$w_2 \neq firma$	$O_{21} = 526$	$O_{22} = 1608835$

Tabulka 5.9: Kontingenční tabulka použitá v χ^2 testu při odhalování závislosti mezi výskyty slova *každá* a *firma*. V textu je 6 výskytů tohoto bigramu, 8 478 bigramů, jejichž druhé slovo je *firma*, ale první slovo není *každá*, 526 bigramů, u kterých je první slovo *každá* a druhé slovo není *firma*, a konečně 1 608 835 bigramů, které neobsahují ani jedno z těchto dvou slov.

a rozptylu:

$$s^2 \approx p' \approx 5,56296 \times 10^{-6}.$$

Nyní máme všechny údaje nutné pro výpočet hodnoty T:

$$T = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5,56296 \times 10^{-6} - 1,55327 \times 10^{-6}}{\sqrt{\frac{5,56296 \times 10^{-6}}{1617845}}} \approx 2,1624$$

Výsledná hodnota $T = 2,1624$ je menší než $2,576$, což je příslušná kritická hodnota pro $\alpha = 0,005$, proto nulovou hypotézu o nezávislosti výskytu slov *nová* a *společnost* zamítnout nemůžeme. Značí to tedy, že fráze *nová společnost* je plně kompoziční, nevytváří žádný nový význam a není kolokací (s pravděpodobností 99,5 %).

Tabulka 5.7 obsahuje hodnoty statistiky T za předpokladu platnosti H_0 pro deset bigramů z kolekce $\mathcal{CW}94$, které se v něm vyskytují právě dvacekrát. U prvních pěti můžeme nulovou hypotézu na hladině $1 - \alpha$ ($\alpha = 0,005$) zamítnout, a proto jsou vhodnými kandidáty na kolokace. U zbylých pěti bigramů hypotézu zamítnout nelze, prokázala se tedy nezávislost jednotlivých komponent bigramů, a proto je jako vhodné kandidáty na kolokace označit nelze.

Všimněme si, že metoda založená na počítání výskytů bigramů v textu z kapitoly 5.1 by na dvojicích z tabulky 5.7 ztroskotala. Všechny bigramy mají stejnou frekvenci výskytu a tato metoda by je nedokázala rozlišit. Její výhodou je, že bere v úvahu i frekvence samostatných slov.

5.3.2 χ^2 test

t test má jeden důležitý nedostatek. Počítá totiž s tím, že pravděpodobnosti výskytu slov mají přibližně normální rozdělení, což není obecně pravda([13]).

Alternativní test, který je také určen pro zjišťování závislostí jevů, ale nepředpokládá normalitu rozdělení, je χ^2 *test* (označovaný také jako Pearsonův).

Při aplikaci tohoto testu se používá tzv. *kontingenčních tabulek*. V nejjednodušším případě, kdy testujeme závislost dvou jevů (výskyt slova w_1 a výskyt slova w_2), je kontingenční tabulka dvojrozměrná, nebo také tzv. *čtyřpolní* [12]. Viz tabulka 5.8. Pokud by předmětem našeho zkoumání byly trigramy a my bychom zjišťovali závislosti výskytů tří slov, kontingenční tabulka by byla trojrozměrná a měla by 8 polí.

Kontingenční tabulku pro bigram w_1w_2 vytvoříme takto: Analyzovaný text bereme opět jako posloupnost bigramů a zkoumáme shodu prvního slova bigramu se slovem w_1 a shodu druhého slova bigramu se slovem w_2. V závislosti na tomto pozorování může nastat jedna ze 4 situací: a) první slovo je w_1 a druhé slovo je w_2, b) první slovo není w_1 a druhé slovo je w_2, c) první slovo není w_1 a druhé slovo je w_2 a konečně d) první slovo není w_1 a ani druhé slovo není w_2. Nás zajímá, kolikrát ta která situace nastala, a odpovídající četnosti jsou právě prvky kontingenční tabulky (příklad viz tabulka 5.9).

χ^2 test porovnává empiricky zjištěné hodnoty kontingenční tabulky s hodnotami, které lze očekávat v případě, že platí nulová hypotéza, která říká, že pozorované výskyty slov jsou nezávislé. Pokud je rozdíl naměřených a očekávaných hodnot velký, nulovou hypotézu zamítáme.

V tabulce 5.9 jsou četnosti bigramů, které obsahují slova *každá* a *firma* v kolekci $\mathcal{CW}94$. Frekvence těchto slov jsou následující:

$$\begin{aligned} C(\textit{každá}) &= 8484 \\ C(\textit{firma}) &= 532 \\ C(\textit{každá firma}) &= 6 \\ N &= 1617845 \end{aligned}$$

Z výše uvedeného vyplývá, že počet bigramů w_iw_{i+1} takových, že první slovo není *každá* a druhé slovo je *firma*, je:

$$C(\neg \textit{každá}, \textit{firma}) = 8484 - 6 = 8478$$

A analogicky:

$$\begin{aligned} C(\textit{každá}, \neg \textit{firma}) &= 532 - 6 = 526 \\ C(\neg \textit{každá}, \neg \textit{firma}) &= 1617845 - 532 - 8484 + 6 = 1608835 \end{aligned}$$

Nyní se dostáváme k samotnému výpočtu hodnoty χ^2. Získá se jako suma čtverců rozdílu naměřených a očekávaných hodnot z tabulky 5.9 vážených velikostí (významem) očekávaných hodnot.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5.6}$$

i reprezentuje řádky tabulky, j sloupce; O_{ij} jsou naměřené hodnoty a E_{ij} hodnoty očekávané. Je možné dokázat, že veličina χ^2 má asymptoticky (při dostatečně velkých výběrech) χ^2_1 rozdělení (stupeň volnosti 1 pro pozorování závislosti dvou jevů).

Očekávané hodnoty E_ij zjistíme ze znalostí naměřených hodnot O_{ij} a počtu všech bigramů v textu N. Postup pro zjištění hodnoty E_{11} je následující. Víme:

$$\begin{aligned} P(w_1) &= \frac{C(w_1)}{N} = \frac{O_{11}+O_{12}}{N} \\ P(w_2) &= \frac{C(w_2)}{N} = \frac{O_{11}+O_{21}}{N} \end{aligned}$$

Předpokládáme, že platí nulová hypotéza:

$$H_0 : P(w_1 w_2) = P(w_1)P(w_2)$$

A dále:

$$\begin{aligned} E_{11} &= C(w_1 w_2) \\ &= P(w_1 w_2) \times N \\ &= P(w_1)P(w_2) \times N \\ &= \frac{C(w_1)}{N} \times \frac{C(w_2)}{N} \times N \\ &= \frac{O_{11}+O_{12}}{N} \times \frac{O_{11}+O_{21}}{N} \times N \end{aligned}$$

Analogicky:

$$\begin{aligned} E_{12} &= \frac{O_{11}+O_{12}}{N} \times \frac{O_{22}+O_{12}}{N} \times N \\ E_{21} &= \frac{O_{21}+O_{22}}{N} \times \frac{O_{11}+O_{21}}{N} \times N \\ E_{22} &= \frac{O_{22}+O_{12}}{N} \times \frac{O_{22}+O_{21}}{N} \times N \end{aligned}$$

V případě použití dvojrozměrné tabulky lze výpočet χ^2 zjednodušit do této podoby:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \tag{5.7}$$

Pro tabulku 5.9 je hodnota χ^2 získaná tímto vztahem následující:

$$\chi^2 = \frac{1617845(6 \times 1608835 - 526 \times 8478)^2}{(6+8478)(6+526)(8478+1608835)(526+1608835)} \approx 3,716$$

Ve statistických tabulkách lze nalézt kritickou hodnotu χ^2_1 rozdělení rovnou $3,841$ ($\alpha = 0,05$ a jeden stupeň volnosti pro zkoumání závislosti dvou jevů). Protože $3,716 < 3,841$ nulovou

χ^2	$C(w_1)$	$C(w_2)$	$C(w_1, w_2)$	w_1	w_2
1 035 412,7998	25	25	20	červený	kříž
330 156,5304	20	98	20	řádová	čárka
417 489,2901	50	31	20	San	Francisko
29 452,7674	32	686	20	aritmetická	operace
23 055,2840	77	364	20	podavač	papíru
2,8645	9 781	2 274	20	systém	typu
2,8777	21 394	1 040	20	na	další
1,2213	5 300	4 775	20	program	v
1,2357	769	32 922	20	technika	je
0,6451	823	32 922	20	úroveň	je

Tabulka 5.10: Výsledky aplikace χ^2 *testu* na bigramy z kolekce $\mathcal{CW}94$ s frekvencí právě 20

hypotézu, že *nová* a *společnost* se v textu vyskytují nezávisle na sobě, zamítnout nemůžeme a bigram *nová společnost* tedy není vhodný kandidát na kolokaci.

Tohoto výsledku jsme dosáhli také při použití *t testu* a nabízí se tvrzení, že rozdíly v *t testu* a χ^2 *testu* pro rozpoznávání kolokací nejsou velké. χ^2 *test* je podle Christophera Manninga [6] vhodný v případech, kdy mají studované jevy větší pravděpodobnost výskytu a *t test* v nich tolik úspěšný není. Použití *t testu* je problematické předpokladem normálního rozdělení. Podobné problémy nastávají při aplikaci χ^2 *testu* v případech, kdy jsou hodnoty v tabulce 5.9 malé. Snedecor a Cochran [14] nedoporučují používat χ^2 *test* v případech, kdy je celková velikost vzorku menší než 20, nebo pokud je menší než 40 a očekávané hodnoty v tabulce jsou menší nebo rovny 5.

5.3.3 Poměr pravděpodobností

Další metoda testování hypotéz plyne z měření tzv. *poměru pravděpodobností*. Oproti předchozím testům má dvě výhody. Lze ji použít v případech s menšími frekvencemi slov a lépe se interpretuje její výsledek. Jednoduše říká, o kolik je jedna hypotéza pravděpodobnější než jiná.

Při aplikaci této metody pro rozpoznávání kolokací se zkoumají dvě alternativní hypotézy o pravděpodobnostech výskytu bigramu w_1w_2 [15]:

$$H_1 : \quad P(w_2|w_1) = p = P(w_2|\neg w_1) \tag{5.8}$$

$$H_2 : \quad P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1) \tag{5.9}$$

Hypotéza H_1 formalizuje nezávislost výskytů slov w_1 a w_2 podobně jako nulová hypotéza v předchozích testech.

Hypotéza H_2 popisuje závislost výskytů těchto slov. Předpokládáme, že pokud tato hypotéza platí, je $p_1 \gg p_2$. Případy, kdy $p_1 \ll p_2$, jsou ojedinělé a budeme je zanedbávat.

Pomocí četnosti výskytů slov w_1, w_2 a bigramu w_1w_2 v textech spočteme maximální věrohodné odhady p, p_1 a p_2:

	H_1	H_2
$P(w_2\|w_1)$	$p = \frac{c_2}{N}$	$p_1 = \frac{c_{12}}{c_1}$
$P(w_2\|\neg w_1)$	$p = \frac{c_2}{N}$	$p_2 = \frac{c_2 - c_{12}}{N - c_1}$
$P(c_{12}$ bigramů z c_1 jsou $w_1w_2)$	$b(c_{12}; c_1, p)$	$b(c_{12}; c_1, p)$
$P(c_2 - c_{12}$ bigramů z $N - c_1$ jsou $\neg w_1 w_2)$	$b(c_2 - c_{12}; N - c_1, p)$	$b(c_2 - c_{12}; N - c_1, p_2)$

Tabulka 5.11: Výpočet testu poměrem pravděpodobností. Pravděpodobnost hypotézy je vždy součin posledních dvou řádků v odpovídajícím sloupci.

$$\begin{array}{ll} c_1 = C(w_1) & p = \frac{c_2}{N} \\ c_2 = C(w_2) & p_1 = \frac{c_{12}}{c_1} \\ c_{12} = C(w_1w_2) & p_2 = \frac{c_2 - c_{12}}{N - c_1} \end{array}$$

Při výpočtu použijeme binomické rozdělení:

$$b(k; n, x) = \binom{n}{k} x^k (1-x)^{(n-k)}$$

a pro obě hypotézy stanovíme pravděpodobnosti, že očekávané frekvence slov w_1, w_2 a bigramu w_1w_2 budou takové, jaké jsme naměřili. Diskuse, jak tyto pravděpodobnosti získáme, je v tabulce 5.11. Pro hypotézu H_1 je výpočet následující:

$$L(H_1) = b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p) \tag{5.10}$$

a pro hypotézu H_2 tento:

$$L(H_2) = b(c_{12}; c_1, p_1) b(c_2 - c_{12}; N - c_1, p_2) \tag{5.11}$$

Poměr pravděpodobností hypotéz λ je potom:

$$\lambda = \frac{L(H_1)}{L(H_2)} = \frac{b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p)}{b(c_{12}; c_1, p_1) b(c_2 - c_{12}; N - c_1, p_2)} \tag{5.12}$$

V úvodu kapitoly bylo uvedeno, že poměr pravděpodobností lze velmi jednoduše interpretovat. Např. pro bigram *informační středisko* je hodnota $\lambda = e^{0,5 \times 453,6} = 3,1476 \times 10^{98}$ a tak pro něj platí, že je $3,1476 \times 10^{98}$ *krát* pravděpodobnější, že výskyt slova *informační* je závislý na výskytu slova *středisko*, než že mezi nimi žádná závislost není. Jak víme, tato závislost je dobrým indikátorem, zda je bigram kolokací. Čím je větší hodnota λ, tím větší je pravděpodobnost závislosti a větší šance, že se jedná o kolokaci.

Další výhodou testu poměru pravděpodobností je to, že jej lze použít s větším úspěchem než χ^2 *test*, a to v případě řídkých dat, tzn. při malých frekvencích w_1, w_2 nebo w_1w_2. Poměr pravděpodobností hypotéz lze využít i při klasickém testování hypotéz.

Veličina χ^2 má totiž asymptoticky χ^2_2 rozdělení ([6] a [16]) a její hodnoty můžeme použít pro testování nulové hypotézy H_1 proti alternativní hypotéze H_2 obdobně jako v případě χ^2

$-2log\lambda$	$C(w_1)$	$C(w_2)$	$C(w_1,w_2)$	w_1	w_2
7 757,56	639	9 781	567	informační	systém
1 540,57	236	1 607	100	informační	technologie
542,43	236	1 351	42	informační	služba
453,60	108	400	26	informační	středisko
176,05	108	255	11	informační	centrum
73,13	639	32	5	informační	bulletin
68,38	236	4 547	12	informační	síť
33,34	236	1 314	5	informační	hodnota
17,69	236	6	1	informační	tabule
14,04	639	170	2	informační	materiál

Tabulka 5.12: Výsledky aplikace χ^2 *testu* na bigramy z kolekce $\mathcal{CW}94$ s frekvencí právě 20

testu. Kritická hodnota pro $-2log\lambda$ na hladině spolehlivosti 99,5 % ($\alpha = 0,005$) s jedním stupněm volnosti je $7,88$.

Výpočet přirozeného logaritmu λ je následující:

$$\begin{aligned} log\lambda &= log\frac{L(H_1)}{L(H_2)} \\ &= log\frac{b(c_{12};c1,p)b(c_2-c_{12};N-c_1,p)}{b(c_{12};c1,p_1)b(c_2-c_{12};N-c_1,p_2)} \\ &= logL(c_{12},c_1,p)+logL(c_2-c_{12},N-c_1,p) \\ &\quad -logL(c_{12},c_1,p_1)-logL(c_2-c_{12},N-c_1,p_2), \end{aligned}$$

kde $L(k,n,x)=x^k(1-x)^{n-k}$.

Příklady bigramů a odpovídajících hodnot $-2log\lambda$ včetně frekvencí jsou v tabulce 5.12. Všimněme si, že i pro bigramy s relativně malou frekvencí jsou výsledky dobré. Např. *informační tabule* nebo *informační materiál*.

5.4 Vzájemná informace

Další metodou detekce kolokací v textu, která tentokrát vychází z teorie informace, je měření tzv. *vzájemné informace* ([4], [17]). Původní definice zavedená Fanem v roce 1961 ([18]) se podstatně liší od definice používané dnes. Fano definoval vzájemnou informaci mezi dvěma konkrétními událostmi x a y, v našem případě výskyty dvou konkrétních slov, následovně:

$$I(x,y) = log_2\frac{P(xy)}{P(x)P(y)} \tag{5.13}$$

$$= log_2\frac{P(x|y)}{P(x)} \tag{5.14}$$

$$= log_2\frac{P(x|y)}{P(y)} \tag{5.15}$$

$I(w_1, w_2)$	$C(w_1)$	$C(w_2)$	$C(w_1, w_2)$	w_1	w_2
15,6598	25	25	20	červený	kříž
14,0109	20	98	20	řádová	čárka
14,3495	50	31	20	San	Francisko
10,5255	32	686	20	aritmetická	operace
10,1730	77	364	20	podavač	papíru
0,5407	9 781	2 274	20	systém	typu
0,5402	21 394	1 040	20	na	další
0,3545	5 300	4 775	20	program	v
0,3539	769	32 922	20	technika	být
0,2560	823	32 922	20	úroveň	být

Tabulka 5.13: Hodnoty vzájemné informace pro bigramy z kolekce $\mathcal{CW}94$ s frekvencí právě 20

Zjednodušeně můžeme říct, že tento typ vzájemné informace je měřítkem, kolik informace nám výskyt jednoho slova říká o výskytu slova druhého.

V teorii informace se dnes častěji definuje vzájemná informace jako vztah mezi *náhodnými veličinami*, tedy nikoli mezi *hodnotami náhodných veličin*. Odpovídá očekávané hodnotě (*expectation*) předchozího výrazu:

$$I(X;Y) = E_{p(x,y)} log_2 \frac{P(X,Y)}{P(X)P(Y)} \tag{5.16}$$

V komputační lingvistice a NLP se používá původní definice zavedená Fanem. Dnes se pro takto definovanou vzájemnou informaci používá označení *bodová vzájemná informace*. Pro rozpoznávání kolokací ji budeme používat i my.

Odhady pravděpodobnosti nutné pro výpočet spočteme opět jako maximálně věrohodné odhady.

Příklad: Výpočet vzájemné informace slov *San* a *Francisko*

$$I(San\ Francisko) = log_2 \frac{\frac{50}{1617845}}{\frac{31}{1617845} \times \frac{20}{1617845}} = 14,3495$$

V tabulce 5.13 jsou hodnoty vzájemné informace pro bigramy z tabulky 5.7. Jejich klasifikace je stejná jako při použití *t testu*.

Co přesně hodnota vzájemné informace $I(x, y)$ vyjadřuje? Fano o své definici ([18]) píše:

> Množství informace poskytované výskytem události y o výskytu události x je definováno jako:
>
> $$I(x, y) = log_2 \frac{P(x|y)}{P(x)}$$

Například nám hodnota vzájemné informace říká, že množství informace, které máme o výskytu slova *San* na pozici i v textu se zvýší o 14.3495 bitů, pokud se dozvíme, že na pozici $i + 1$ se vyskytuje slovo *Francisko*. Ekvivalentně platí, že množství informace, které máme

	chambre	¬ *chambre*	I	χ^2
house	31 950	12 004		
¬ *house*	4 793	848 330	4,1	553 610
	communes	¬ *communes*		
house	4 974	38 980		
¬ *house*	441	852 682	4,2	88 405

Tabulka 5.14: Korespondence slov *chambre*, *house* a *communes* v korpusu Hanshard. Pomocí χ^2 *testu* a vzájemné informace se snažíme odhalit správný překlad slova *house*. Zatímco vzájemná informace má vyšší skóre u dvojice (*communes, house*), z čehož vyplývá, že hledaný překlad je spíše *communes*, správný překlad odahalí χ^2 *test* jako (*chambre*).

o výskytu slova *Francisko* na pozici $i + 1$ v textu, se zvýší o 14.3495 bitů, pokud se dozvíme, že na pozici i se vyskytuje slovo *San*.

Jiný výklad je tento: Vzájemná informace měří snížení nejistoty o výskytu jednoho slova, když máme informaci o výskytu slova jiného. Můžeme také říct, že pokud víme, že se na pozici i vyskytuje slovo *San*, tak se naše nejistota předpovědi následujícího slova na pozici $i + 1$ sníží o 14.3495 bitů. Jinými slovy, pokud víme, že aktuální slovo je *San*, pak slovo následující bude velmi pravděpodobně *Francisko*.

V mnoha případech však měření této „zvětšené informace“ není dobrým měřítkem závislosti dvou jevů. Tuto skutečnost již dokázali mnozí autoři (viz [6]). Uvedeme příklad, kdy aplikace χ^2 *testu* vede na rozdíl od měření vzájemné informace k lepším, resp. správným výsledkům.

Jedním z prvních použití χ^2 *testu* bylo hledání odpovídajících si dvojic slov v anglickém a francouzském jazyce v *Hanshardu*. Hanshard je korpus textů rozprav kanadského parlamentu, jehož všechny věty jsou zaznamenány jak v angličtině, tak i francouzštině (tzv. *paralelní korpus*). Pro každou větu se sestavily všechny možné dvojice slov, první slovo z anglické věty, druhé z francouzské a spočetly se jejich frekvence v celém korpusu. Výpočtem a porovnáním hodnot statistiky χ^2 se zjišťovaly správné překladové dvojice.

Případ, kdy použití vzájemné informace nedává dobré výsledky, ukážeme na hledání francouzského ekvivalentu anglického slova *house*. Toto slovo se velmi často objevuje v překladech francouzských vět, které obsahují slova *chambre* a *communes*. Je to z toho důvodu, že nejčastější použití slova *house* je *House of Commons*, což odpovídá *Chambre de communes* ve francouzštině. A my se ptáme, jaký je překlad anglického *house*: *chambre* nebo *communes*? Výsledky vidíme v tabulce 5.14. Zatímco χ^2 *test* správně určí *chambre*, vzájemná informace ukazuje na *communes*.

Proč tomu tak je, ukážeme na výpočtu $I(house, chambre)$ a $I(house, communes)$:

$$log\frac{P(house|chambre)}{P(house)} = log\frac{\frac{31950}{31950+4793}}{P(house)} = log\frac{0,87}{P(house)}$$

$$< log\frac{0,92}{P(house)} = log\frac{\frac{4974}{4974+441}}{P(house)} = log\frac{P(house|communes)}{P(house)}$$

Výskyt francouzského *communes* s anglickým *house* v ekvivalentních větách je pravděpodobnější než výskyt *chambre* s *house*. Větší hodnota vzájemné informace pro (*house, communes*) je

I_{50}	w_1	w_2	w_1w_2	$bigram$	I	w_1	w_2	w_1w_2	$bigram$
15,52	1	1	1	neurotický samotář	19,63	1	2	1	neurotický samotář
13,20	1	5	1	umožnit opravdu	12,33	1	315	1	kryptografická ochrana
12,94	1	6	1	kryptografická ochrana	12,58	265	1	1	trochu pomačkat
12,52	8	1	1	trochu pomačkat	10,24	6	223	1	umožnit opravdu
10,20	8	5	1	dobrý vzhled	7,41	160	1 067	18	nižší cena
9,20	8	10	1	nižší cena	6,28	200	104	1	dobrý vzhled
7,35	18	16	1	svůj druhý	4,95	469	333	3	svůj druhý
6,99	23	16	1	obsahovat číslo	4,69	2 440	384	15	však tvrdit
4,78	66	26	1	však tvrdit	1,11	860	869	1	obsahovat číslo
3,26	19	265	1	provoz systému	0,94	688	9 781	8	provoz systému

Tabulka 5.15: Ukázka jednoho z problémů vzájemné informace. V tabulce jsou hodnoty vzájemné informace pro namátkou vybrané bigramy z kolekce 50 dokumentů z $\mathcal{CW}94$ (vlevo) a hodnoty vzájemné informace pro stejné bigramy spočtené pro celou kolekci textů $\mathcal{CW}94$ (vpravo). Tento příklad ilustruje, jak umělé zvýšení hodnoty vzájemné informace pro málo frekventované bigramy, tak zkreslení odhadů pravděpodobností výskytu jednotlivých slov. Zatímco v prvním měření byl maximální věrohodný odhad pravděpodobnosti výskytu bigramu *umožnit opravdu* $P_{50} = \frac{C_{50}}{N_{50}} = \frac{1}{34808}$, v druhém případě je $P = \frac{C}{N} = \frac{1}{1255385}$, tedy asi 35 krát menší.

zapříčiněna tím, že *communes* znamená větší snížení nejistoty než *chambre*. Vidíme tedy, že stupeň nejistoty nekoresponduje s tím, co chceme měřit. Naopak χ^2 *test* je přímým testem závislosti, což svědčí o zajímavém vztahu mezi pozorovanými slovy. At'už se jedná o známku dobrého překladu (jako v tomto příkladě) nebo náznak kolokace (což je důležitější pro nás).

V tabulce 5.15 vidíme další problém s použitím vzájemné informace pro hledání kolokací. Jsou v ní hodnoty vzájemné informace pro bigramy, které se v prvních 50 dokumentech kolekce $\mathcal{CW}94$ vyskytují právě jednou, a hodnoty vzájemné informace stejných bigramů spočítané již pro kompletní kolekci. Nejprve si všimněme, že pro tak málo frekventované dvojice slov jsou hodnoty vzájemné informace poměrně vysoké, aniž by to bylo znakem kolokace. Měření na celé kolekci přineslo velmi malé zlepšení výsledků. I když se měření ale provádělo na datech více než 20 krát větších, 6 bigramů má stále jen jeden výskyt, tím pádem nepřesný maximální věrohodný odhad pravděpodobnosti a také uměle zvýšenou hodnotu MI. Většina z nich nejsou kolokace a bylo by dobré je z množiny potenciálních kolokací odstranit.

Metoda vzájemné informace je poměrně spolehlivá metoda pro měření nezávislosti. Hodnoty kolem 0 implikují nezávislost výskytu jednotlivých slov. Pro měření závislosti však příliš vhodná není, neboť hodnoty vzájemné informace závisí na frekvenci jednotlivých slov. Pro bigramy, jejichž komponenty jsou v textu řádově stejně časté, lišící se jen počtem svých výskytů, platí, že ten méně častý bude mít větší hodnotu MI než ten frekventovanější. Toto je vlastnost, kterou chceme potlačit. Jedno řešení je rovnou vynechat všechny bigramy s četností výskytů menší než např. 3. Problém to sice nevyřeší, ale alespoň eliminuje nejvíce zkreslené výsledky.

5.5 Syntaktická závislost

Doposud jsme příklady použití jednotlivých metod ukazovali na bigramech, které jsme získali z přirozené posloupnosti slov v textu. Ve většině případů byly bigramy těsně sousedící slova v povrchovém slovosledu. Pouze u metody střední hodnoty a rozptylu jsme brali v úvahu i dvojice vzdálenějších slov. Všechny popsané metody lze použít i na bigramy získané ze stromu syntaktických závislostí jednotlivých vět, jak bylo popsáno v kapitole 4.4. V následujících kapitolách si ukážeme, jakým způsobem lze rozšířit testování hypotéz a měření vzájemné informace, pokud máme v anotovaných textech k dispozici také hodnoty analytické funkce.

Nejdříve ale diskutujme smysl použití analytické funkce k tomuto účelu. V kapitole 4.4 jsme se zabývali různými způsoby získávání bigramů jako potenciálních kolokací. Šlo nám především o dvě věci: a) Rozpoznat všechny výskyty konkrétního bigramu v textu ve všech jeho možných tvarech, tzn. „na některé výskyty bigramu nezapomenout“. Toho lze docílit tím, že pracujeme se základními tvary slov, které jsme definovali tamtéž. b) Počítat jen a pouze ty výskyty bigramů, které opravdu jsou zkoumanou jednotkou, tzn. vyhnout se situaci, kdy nějaká dvojici slov bude označena za výskyt konkrétní fráze a ona touto frází nebude. Tento případ spolehlivě vyřešíme, pokud budeme bigramy brát jako slova, mezi kterými je syntaktická závislost, ještě lépe, pokud tuto relaci budeme brát jako nesymetrickou a ještě lépe, pokud budeme brát v úvahu i hodnotu příslušné analytické funkce.

5.5.1 Vzájemná informace

Následující výpočty vzájemné informace syntakticky závislých slov v anglických textech nastínil v roce 1998 Dekang Lin [17]. Tato metoda pracuje s tzv. *závislostními trojicemi*, které se skládají z *řídícího slova*, *typu závislosti* a *slova závislého*. Řídící a závislé slovo jsou v základním tvaru definovaném v kapitole 4.1. Typ závislosti se skládá z slovního druhu řídícího slova, hodnoty analytické funkce a slovního druhu slova závislého (jednotlivé části oddělujeme dvojtečkou). Příklad vytvoření závislostních dvojic vidíme na obrázku 5.3. Popis značení slovních druhů a hodnot analytické funkce jsou v příloze.

Závislostní trojici (w_1, rel, w_2) můžeme brát jako společný výskyt tří jevů:

A: náhodně vybrané slovo je w_1

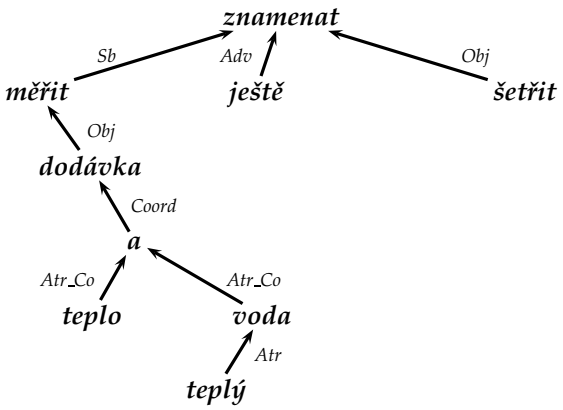
B: náhodně vybraná závislost je rel

C: náhodně vybrané slovo je w_2

Vzájemná informace pro trojici jevů je definována následovně [19]:

$$I(A,B,C) = log_2 \frac{P(A,B,C)}{P(A) \times P(B) \times P(B)} \tag{5.17}$$

Definice předpokládá, že pokud závislostní trojice nereprezentuje kolokaci, jsou výskyty jevů *A, B, C* na sobě nezávislé. To však není pravda. Tyto jevy do jisté míry závislé jsou. Typ závislosti rel určují slovní druhy řídícího a závislého slova. Např. $rel = N : Atr : A$ implikuje, že slovní druh řídícího slova je podstatné jméno (N) a slovní druh závislého slova je přídavné jméno (A).




```
(neznamenat V:Sb:V měřit)    (a J:Attr_Co:N voda)
(měřit V:Obj:N dodávka)      (voda N:Atr:A teplá)
(dodávka N:Coord:J a)        (neznamenat V:Adv: ještě)
(a J:Attr_Co:A teplá)        (neznamenat V:Obj: šetřit)
```

Obrázek 5.3: Sestavení závislostních trojic ve větě: *Měřit dodávky tepla a teplé vody ještě neznamená šetřit.*

Obrázek 5.4: Dva modely Bayesovských sítí reprezentující závislostí jevů *A*, *B*, *C*: a) nezávislost, b) závislost B → A a B→C

Viz příklad *(voda N:Atr:A teplá)*. Na obrázku 5.4 jsou dva modely závislosti jevů *A*, *B*, *C*. Definice vzájemné informace (5.17) odpovídá modelu a). Reálnější je však model b). A pro něj Lin [17] doporučuje následující výpočet vzájemné informace:

$$I(A,B,C) = log_2 \frac{P(A,B,C)}{P(B) \times P(A|B) \times P(C|B)} \tag{5.18}$$

Maximální věrohodné odhady pravděpodobností získáme takto:

$$\begin{aligned} P(A,B,C) &= \frac{C(A,B,C)}{N} \\ P(B) &= \frac{C(*,B,*)}{N} \\ P(A|B) &= \frac{C(A,B,*)}{C(*,B,*)} \\ P(C|B) &= \frac{C(*,B,C)}{C(*,B,*)} \end{aligned}$$

I	I_{rel}	C_{rel}	C	rel	w_1w_2
14,16	12,52	10	10	N:Atr:A	osvědčení lustrační
13,67	12,00	10	10	N:Atr:A	opona železná
11,18	9,44	10	10	V:Obj:N	utrpět zranění
8,61	7,31	10	10	N:Atr:A	výbor rozpočtový
7,21	7,89	10	10	N:Atr:N	novelizace zákon
4,73	2,21	10	14	N:Adv:N	průběh roku
4,41	1,88	10	10	N:Atr:C	týden první
4,25	4,15	10	10	N:Atr:N	cena zboží
4,07	2,86	10	10	N:Atr:N	většina země
2,93	2,31	10	10	N:Atr:N	řada firem

Tabulka 5.16: Hodnoty vzájemné informace pro bigramy s četností právě 10 a pro odpovídající závislostní trojice (z korpusu $\mathcal{PDT}$). I_{rel} je vzájemná informace získaná rozšířením bigramu na závislostní trojici, I původní vzájemná informace, C_{rel} je četnost závislostní trojice, C četnosti bigramu, rel typ závislosti a w_1w_2 zkoumaná fráze.

kde N je počet všech závislostních trojic, $C(A, B, C)$ je počet závislostních trojic s jevy (A, B, C). Pokud se na některém místě trojice vyskytuje znak $*$, znamená to, že se počet trojic sčítá přes všechna možná slova, resp. relace. Např. $(*, rel, *)$ odpovídá všem závislostním trojicím, ve kterých je typ závislosti shodný s rel.

Výsledky rozšíření výpočtu vzájemné informace o typ závislosti jsou obsaženy v tabulce 5.16. Můžeme pozorovat snížení hodnoty MI pro téměř všechny bigramy, ovšem toto snížení je znatelnější u nevhodných kandidátů na kolokace. Čím je kolokace markantnější, tím je snížení menší. V jednom případě došlo dokonce ke zvýšení, jedná se o jasnou kolokaci *novelizace zákona*. Míru kolokativnosti tedy lépe popisuje vzájemná informace definovaná pomocí závislostních trojic. V seznamu dvojslovných frází sestupně setříděném podle nově definované vzájemné informace bude většina kolokací blíže začátku seznamu a naopak nekolokativní bigramy budou na konci.

5.5.2 t test

Stejný princip jako v předchozí kapitole použijeme i pro modifikaci *t testu*. Původní nulová hypotéza zněla takto:

$$H_0 : P(w_1w_2) = P(w_1)P(w_2) \tag{5.19}$$

Předpokládali jsme, že nezávislost výskytu slov w_1 a w_2 je znakem toho, že mezi nimi není žádný zajímavý vztah a že netvoří kolokaci. Nyní bigram w_1w_2 rozšíříme na syntaktickou trojici (w_1, rel, w_2). Vzorek pro testování nebude tedy posloupnost bigramů, ale posloupnost závislostních trojic, z které sestrojíme posloupnost hodnot 0 a 1 analogicky k postupu na obrázku 5.2.

Nabízí se rozšíření nulové hypotézy na tento tvar:

$$H_1 : P(w_1, rel, w_2) = P(w_1)P(rel)P(w_2) \tag{5.20}$$

t	t_{rel}	C_{rel}	C	rel	$w_1 w_2$
3,16	242,45	10	10	N:Atr:A	osvědčení lustrační
3,16	202,83	10	10	N:Atr:A	opona železná
3,16	83,40	10	10	V:Obj:N	utrpět zranění
3,15	39,58	10	10	N:Atr:A	výbor rozpočtový
3,14	38,32	10	10	N:Atr:N	novelizace zákona
3,60	5,33	10	14	N:Adv:N	průběh roku
3,01	4,43	10	10	N:Atr:C	týden první
2,97	7,35	10	10	N:Atr:N	většina země
2,99	12,60	10	10	N:Atr:N	cena zboží
2,74	5,62	10	10	N:Atr:N	řada firem

Tabulka 5.17: Hodnoty statistiky T pro bigramy s četností právě 10 a pro odpovídající závislostní trojice z korpusu $\mathcal{PDT}$. t je hodnota původní statistiky T a t_{rel} modifikované verze, která počítá s typy závislostí.

V předchozí kapitole jsme však zjistili, že tato nezávislost není dobrým znakem „nekolokace“. Mezi výskyty jevů w_1, w_2 a rel totiž závislost existuje a těžko bychom hledali závislostní trojice, pro něž by nebyla hypotéza zamítnuta.

S ohledem na závislost výskytu slov w_1 a w_2 na relaci rel formulujeme novou nulovou hypotézu takto:

$$H_2 : P(w_1, rel, w_2) = P(rel)P(w_1|rel)P(w_2|rel) \tag{5.21}$$

Pokud se nám podaří i tuto hypotézu zamítnout, máme o to vážnější důvod se domnívat, že jsme odhalili kolokaci.

Statistika T zůstane stejná:

$$T = \frac{|\overline{x} - \mu|}{\sqrt{\frac{s^2}{N}}}, \tag{5.22}$$

Změní se pouze způsob výpočtu odhadu střední hodnoty $\overline{x}$, rozptylu s^2 ze vzorku a střední hodnoty Bernouliho rozdělení μ za předpokladu pravdivosti hypotézy H_0:

$$\begin{aligned}
\mu &= P(rel)P(w_1|rel)P(w_2|rel) \\
&= \frac{C(*, rel, *)}{N} \frac{C(w_1, rel, *)}{C(*, rel, *)} \frac{C(*, rel, w_2)}{C(*, rel, *)} \\
\overline{x} &= \frac{C(w_1, rel, w_2)}{N} \\
s^2 &\approx \frac{C(w_1, rel, w_2)}{N}
\end{aligned}$$

Kritická hodnota je stejná jako v případě bigramů $2,576$ ($\alpha = 0,005$). Výsledky testovaní hypotézy H_2 pomocí *t testu* provedené na korpusu $\mathcal{PDT}$ jsou v tabulce 5.17. Vidíme, že počítání s relací syntaktické závislosti vede k velmi znatelnému zvýšení hodnoty statistiky T právě u vhodnějších kandidátů na kolokace.

5.5.3 χ^2 test

Stejným způsobem jako v předešlých případech rozšíříme i χ^2 *test*. V kapitole 5.4 jsme popsali princip aplikace tohoto testu na výskyty dvojic slov w_1, w_2, při kterém se používá dvojrozměrná tabulka naměřených a očekávaných výskytů bigramů obsahující, resp. neobsahující zkoumaná slova (viz. tabulka 5.8). Tím, že bigramy (w_1, w_2) rozšíříme na závislostní trojice (w_1, rel, w_2), zvětší se počet zkoumaných jevů a musíme použít tabulku pro jevy tři, tedy trojrozměrnou. Její schéma je znázorněno v tabulce 5.18.

$O_{111} = \|\|w_1, rel, w_2\|\|$	$O_{121} = \|\|\neg w_1, rel, w_2\|\|$
$O_{211} = \|\|w_1, rel, \neg w_2\|\|$	$O_{221} = \|\|\neg w_1, rel, \neg w_2\|\|$

$O_{112} = \|\|w_1, \neg rel, w_2\|\|$	$O_{122} = \|\|\neg w_1, \neg rel, w_2\|\|$
$O_{212} = \|\|w_1, \neg rel, \neg w_2\|\|$	$O_{222} = \|\|\neg w_1, \neg rel, \neg w_2\|\|$

Tabulka 5.18: Hodnoty z této tabulky se používají v χ^2 *testu* při odhalování závislosti mezi výskyty slov w_1 a w_2 a relace rel. Jedná se o trojrozměrnou variantu tabulky 5.8.

Statistika χ^2 se pak počítá takto:

$$\chi^2 = \sum_{i,j,k} \frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}}, \tag{5.23}$$

kde $i, j, k \in \{1, 2\}$.

Předpokládáme, že platí nulovou hypotéza stejně jako u *t testu* (5.21):

$$H_2 : P(w_1, rel, w_2) = P(rel)P(w_1|rel)P(w_2|rel)$$

Empiricky zjištěné hodnoty tabulky 5.18 získáme z frekvencí jednotlivých jevů:

$$\begin{aligned}
O_{111} &= C(w_1, rel, w_2) \\
O_{121} &= C(*, rel, w_2) - C(w_1, rel, w_2) \\
O_{211} &= C(w_1, rel, *) - C(w_1, rel, w_2) \\
O_{221} &= C(*, rel, *) - C(w_1, rel, *) - C(*, rel, w_2) + C(w_1, rel, w_2) \\
O_{112} &= C(w_1, *, w_2) - C(w_1, rel, w_2) \\
O_{122} &= C(*, *, w_2) - C(*, rel, w_2) - C(w_1, *, w_2) + C(w_1, rel, w_2) \\
O_{212} &= C(w_1, *, *) - C(w_1, rel, *) - C(w_1, *, w_2) + C(w_1, rel, w_2) \\
O_{222} &= C(*, *, *) - C(w_1, *, *) - C(*, *, w_2) - C(*, rel, *) + \\
&\quad C(w_1, rel, *) + C(*, rel, w_2) + C(w_1, *, w_2) - C(w_1, rel, w_2)
\end{aligned}$$

Očekávané hodnoty tabulky 5.18 za předpokladu platnosti nulové hypotézy H_2:

$$E_{111} = N \times P(rel)P(w_1|rel)P(w_2|rel)$$

$$
\begin{aligned}
E_{121} &= N \times P(rel)P(\neg w_1|rel)P(w_2|rel) \\
E_{211} &= N \times P(rel)P(w_1|rel)P(\neg w_2|rel) \\
E_{221} &= N \times P(rel)P(\neg w_1|rel)P(\neg w_2|rel) \\
E_{112} &= N \times P(\neg rel)P(w_1|\neg rel)P(w_2|\neg rel) \\
E_{122} &= N \times P(\neg rel)P(\neg w_1|\neg rel)P(w_2|\neg rel) \\
E_{212} &= N \times P(\neg rel)P(w_1|\neg rel)P(\neg w_2|\neg rel) \\
E_{222} &= N \times P(\neg rel)P(\neg w_1|\neg rel)P(\neg w_2|\neg rel)
\end{aligned}
$$

Použité pravděpodobnosti určíme opět metodou maximální věrohodnosti z frekvencí příslušných trojic:

$$
\begin{aligned}
N &= C(*,*,*) \\
P(rel) &= \frac{C(*,rel,*)}{N} \\
P(\neg rel) &= 1 - \frac{C(*,rel,*)}{N} \\
P(w_1|rel) &= \frac{C(w_1,rel,*)}{C(*,rel,*)} \\
P(w_2|rel) &= \frac{C(*,rel,w_2)}{C(*,rel,*)} \\
P(\neg w_1|rel) &= 1 - \frac{C(w_1,rel,*)}{C(*,rel,*)} \\
P(\neg w_2|rel) &= 1 - \frac{C(*,rel,w_2)}{C(*,rel,*)} \\
P(w_1|\neg rel) &= \frac{C(w_1,*,*) - C(w_1,rel,*)}{N - C(*,rel,*)} \\
P(w_2|\neg rel) &= \frac{C(*,*,w_2) - C(*,rel,w_2)}{N - C(*,rel,*)} \\
P(\neg w_1|\neg rel) &= 1 - \frac{C(w_1,*,*) - C(w_1,rel,*)}{N - C(*,rel,*)} \\
P(\neg w_2|\neg rel) &= 1 - \frac{C(*,*,w_2) - C(*,rel,w_2)}{N - C(*,rel,*)}
\end{aligned}
$$

Výsledky výpočtu χ^2 *testu* pro hypotézu H_2 obsahuje tabulka 5.19. Opět můžeme pozorovat rozdíly hodnot χ^2 pro bigramy a pro jim odpovídající závislostní trojice. Podobně jako u vzájemné informace jsou hodnoty statistiky χ^2 , která počítá s typy relací, menší. Toto zmenšení je tím znatelnější, čím větší je šance, že se nejedná o kolokaci. Opět vidíme znatelný přínos toho, že pracujeme se závislostními trojicemi a ne pouze s bigramy.

5.6 Filtrace bigramů

V této části se vrátíme ke slovnědruhovému filtru Justensona a Katze zmiňovanému v kapitole 5.1 a ukážeme, jak se dá jednoduchý princip založený na základních lingvistických znalostech rozšířit.

χ^2	χ^2_{rel}	C_{rel}	C	rel	$w_1 w_2$
183 728,84	58 799,99	10	10	N:Atr:A	osvědčení lustrační
131 230,27	41 155,99	10	10	N:Atr:A	opona železná
23 333,27	6 964,92	10	10	V:Obj:N	utrpět zranění
3 894,66	1 570,73	10	10	N:Atr:A	výbor rozpočtový
2 371,88	1 476,42	10	10	N:Atr:N	novelizace zákona
345,01	41,80	10	14	N:Adv:N	průběh roku
149,85	54,57	10	10	N:Atr:N	většina země
57,96	32,26	10	10	N:Atr:N	řada firem
193,85	20,37	10	10	N:Atr:C	týden první
172,12	159,87	10	10	N:Atr:N	cena zboží

Tabulka 5.19: Hodnoty testové statistiky χ^2 pro bigramy s frekvencí právě 20 a pro odpovídající závislostní trojice (z korpusu $\mathcal{PDT}$).

5.6.1 Slovní druhy

Justensonův a Katzeův filtr je velmi úspěšný. Je to tím, že se soustředí na kolokace nejčastějších slovnědruhových typů. Všechny dvojice slov kromě *přídavné jméno – podstatné jméno* (AN) a *podstatné jméno – podstatné jméno* (NN) zanedbává.

Pro naše potřeby detekce českých kolokací jsou vzory doporučené Justensonem a Katzem až příliš přísné. Nevyhovují jim kolokace obsahující např. slovesa, číslovky nebo příslovce. Christopher Manning a Hinrich Schütze, kteří se také zabývali detekcí dvouslovných kolokací (anglických) [6] doporučují použít seznam stopslov, který vyjme ze zpracování všechna slova, která nejsou podstatnými jmény (N), přídavnými jmény (A) nebo slovesy (V). Po odfiltrování všech ostatních slovních druhů zůstanou v seznamu bigramů pouze dvojice tvořené kombinacemi slovních druhů N, A a V.

V českých textech jsou však také relativně časté i kolokace tvořené číslovkami (C) a příslovci (D). Maximální množina slovnědruhových vzorů pro detekci českých kolokací by tedy měla obsahovat číslovky a příslovce. Tedy všechny *autosémantické slovní druhy*, ne však všechny jejich kombinace. Například vzory CA, CV, CD, AC, VC a DC nemá smysl do zpracování zahrnout, neboť kolokace zpravidla netvoří.

Problematiku slovnědruhových filtrů shrnuje tabulka 5.20. Základní a nejpřísnější filtr Justensona a Katze je v první části. Rozšíření o slovesa dle doporučení Manninga a Schütze je v části druhé a následuje maximální množina slovnědruhových vzorů pro české kolokace v části třetí. Protože ne všechny vzory jsou stejně významné, jsou v tabulce seřazeny podle frekvencí kolokací odpovídajících typů.

5.6.2 Slovní poddruhy

Výsledky morfologické analýzy, popsané v kapitole 4.3, nám umožňují u slovních tvarů rozeznávat i tzv. *slovní poddruhy*. Filtraci můžeme aplikovat i na ně. Ne všechny poddruhy např.

vzor	příklad
A N	operační systém
N N	správce sítě
N A	kód Kamenických
N V	soubor obsahuje
V N	spustit program
V V	jít nakoupit
D A	objektově orientovaný
D V	vzájemně propojit
V D	pracovat společně
A A	stíněný kroucený
C N	devadesátá léta
N D	lícem nahoru
N C	Windows 3.0
D D	velmi dobře
D N	včetně příslušenství
A V	nutno podotknout
C C	první třetina
A D	každý zvlášť

Tabulka 5.20: Doporučené slovnědruhové vzory pro filtrování českých bigramů. V první části jsou vzory doporučené Justensonem a Katzem. Druhá část je rozšíření doporučené Manningem a Schützem a poslední část je naše maximální rozšíření pro české kolokace.

číslovek a sloves bývají součástí kolokací a i jejich odfiltrováním se nám zmenší množina kandidátů na kolokace.

V tabulce 5.21 jsou pro všechny autosémantické slovní druhy uvedené jejich poddruhy, které mohou tvořit kolokace a je vhodné je při detekci kolokací zpracovávat.

5.6.3 Syntaktická závislost

Další možností, jak z množiny zpracovávaných bigramů odstranit ty, které jistě nejsou kolokacemi, je jejich filtrace podle typu syntaktické závislosti mezi řídícím a závislým slovem (čili podle hodnoty analytické funkce), pokud tyto údaje jsou k dispozici. Tabulka 5.22 obsahuje druhy syntaktické závislosti typické pro kolokace.

Další vylepšení jak odfiltrovat nevhodné bigramy spočívá ve zkombinování dvou posledních postupů. V kapitole 2.3 jsme definovali tzv. rozšířené typy syntaktické závislosti. Použili jsme k tomu nejen analytickou funkci, ale i slovní druhy řídícího a závislého slova. Kolokace pak spadají do kategorií v tabulce 5.23.

slovní druh	slovní poddruhy
N	N
A	2 A C G U
C	= a d h l n r v y }
V	B f i p s
D	b g

Tabulka 5.21: Slovní poddruhy, které mohou tvořit kolokace. Vysvětlivky k použitým značkám jsou v příloze.

5.6.4 Sourozenecký syntaktický vztah

Bigramy získané z relace sourozeneckého syntaktického vztahu, kterou jsme definovali v kapitole 4.4, můžeme filtrovat následujícím velmi jednoduchým způsobem. Předpokládáme, že pokud tvoří kolokaci, mají se ke svému řídícímu slovu stejně. Jinými slovy, druhy syntaktické závislosti, které jsou mezi komponentami a společném řídícím slově jsou stejné. Příklady vidíme v tabulce 5.24.

typ závislosti	příklad
Atr	cenný papír
Sb	soud rozhodl
Obj	dávat přednost
Adv	zdravotně postižený

Tabulka 5.22: Hodnoty analytické funkce, které se mohou vyskytovat v kolokacích. Všechny mohou mít „příponu“ *_Co* (součást koordinace), resp. *_Ap* (součást aposice), resp. *_Pa* (řídící uzel vsuvky - parentese).

typ rozš. závislosti	příklad
(V, Sb, N)	dolar posílil
(V, Sb, V)	překvapilo, že prší
(V, Obj, N)	podat hlášení
(V, Obj, V)	přestalo pršet
(A, Obj, N)	podobný barvou
(N, Atr, N)	povrch Země
(N, Atr, A)	zákony přírodní
(V, Adv, D)	sexuálně obtěžovat
(V, Adv, N)	poslat poštou
(A, Adv, D)	silně souvislý
(A, Adv, N)	poháněný akumulátorem
(V, Adv, V)	odejít studovat
(D, Adv, N)	kolmo k přímce

Tabulka 5.23: Typy syntaktické závislosti a s ní spojené slovní druhy použité pro filtraci nevhodných bigramů

typ sour. závislosti	příklad
`Atr:Atr`	bývalý sovětský
`Atr_Co:Atr_Co`	český slovenský
`Sb_Ap:Sb_Ap`	čtenář dopisovatel
`Atr_Ap:Atr_Ap`	organizace OOP

Tabulka 5.24: Bigramy získané ze sourozeneckých syntaktických vztahů lze filtrovat podle typů jejich syntaktické závislostí na společném řídící slově. Velmi přínosné je pracovat jen s těmi bigramy, jejichž komponenty mají stejné druhy syntaktické závislosti na společného předka.

Kapitola 6

Experimenty

V této kapitole se budeme zabývat experimenty souvisejícími s vyhledáváním kolokací v textech a také s jejich hodnocením.

6.1 Výpočty

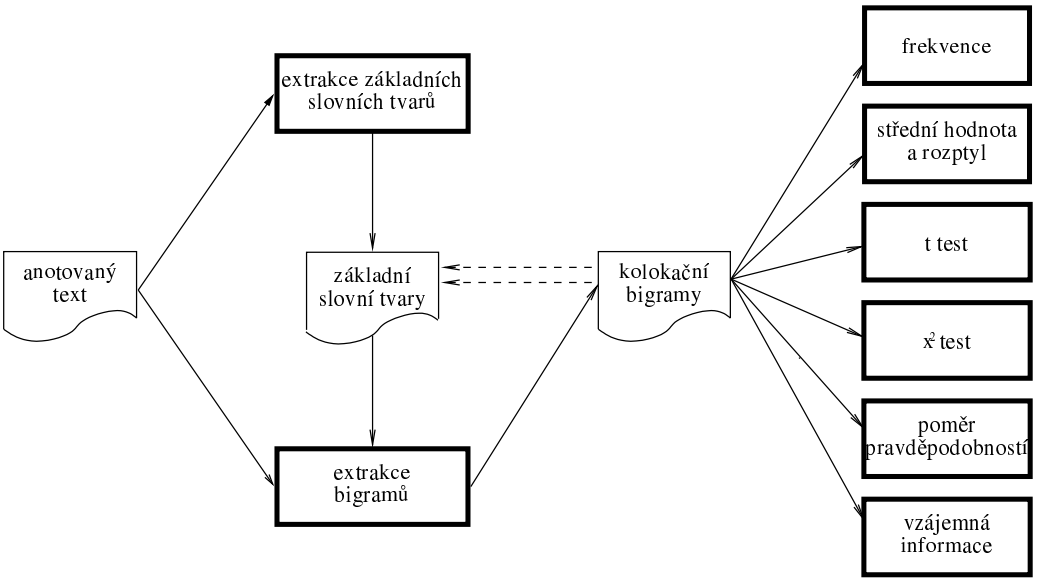
Nyní máme již k dispozici všechny předpoklady pro aplikování metod detekce kolokací v textech. V kapitole 4.4 jsme ukázali, jak lze za pomoci lingvistického parseru zbavit zpracovávanou kolekci nevhodných dokumentů a jak z ní získat kolokující bigramy. Představili jsme celkem šest variant extrakce kolokujících bigramů. Předpokládáme totiž, že způsob získání těchto bigramů ovlivňuje, jaké druhy kolokací se mezi nimi vyskytují.

V kapitole 5 jsme poté prezentovali šest základních metod automatické detekce kolokací. Pro ilustraci jsme tyto metody aplikovali na nepozměněný vstupní text, který jsme považovali za posloupnost bigramů. Každá dvojice sousedících slov v povrchovém slovosledu pro nás byla bigramem, který jsme se snažili ohodnotit podle toho, zda tvoří nebo netvoří kolokaci (resp. do jaké míry kolokaci tvoří nebo netvoří).

My však můžeme tyto metody aplikovat na bigramy získané jakýmkoliv z uvedených způsobů. V následujících kapitolách se budeme zabývat všemi možnými kombinacemi postupů extrakce kolokujících bigramů a metod detekce kolokací v nich.

Všechny experimenty závisí na variantě, kterou jsme získali kolokující bigramy, a mají následující strukturu. Viz obrázek 6.1.

Předpokládáme, že máme připravený anotovaný text zpracovávané kolekce dokumentů. Značkování nám zajistí automatický lingvistický parser [27]. Nejdříve z textu extrahujeme seznam základních tvarů všech slov, která se v něm vyskytují (základní tvar slova - viz kapitola 4.4). Pro každý základní tvar slova zjistíme i počet výskytů všech slov v textu, kterým odpovídá tento základní tvar. V tabulce 6.1 a) je seznam deseti nejčastějších základních tvarů v kolekci $\mathcal{CW}94$. V tabulce 6.1 b) je seznam deseti nejčastějších autosémantických základních slovních tvarů. I nadále budeme, jak jsme zvyklí, místo *základní tvar slova* zkráceně používat *slovo*.



Obrázek 6.1: Schéma průběhu experimentů detekce kolokací. Dvojitá čára vedoucí ze seznamu bigramů do seznamu základních slovních tvarů naznačuje, že všechny bigramy jsou reprezentovány jako dvojice základních slovních tvarů.

V dalším kroku se jedním z popsaných postupů sestaví seznam kolokujících bigramů. Jeho součástí budou údaje především o frekvencích bigramů a v případech, kdy to bude možné, i typy syntaktické závislosti apod.

Takto získaná data se v následujících fázích použijí jako vstupní hodnoty procedur realizujících výpočty jednotlivých metod hledání kolokací. Jejich výsledkem je vždy jedna nebo dvě hodnoty: frekvence, střední hodnota a rozptyl, T statistika, χ^2 statistika, vzájemná informace. Každému bigramu ze seznamu budou přiřazeny hodnoty těchto veličin, které mají korespondovat s tím, zda se jedná nebo nejedná o kolokaci.

Od této chvíle tedy nezpracováváme posloupnost slov, ale posloupnost bigramů. Pro přesnější maximální věrohodné odhady pravděpodobností budeme rozlišovat výskyty slov podle toho, zda se jedná o výskyt na místě prvního (C_1) nebo druhého (C_2) slova bigramu. V případě, že v bigramu rozlišujeme syntaktickou závislost, na prvním místě se uvádí řídící slovo, na druhém slova závislé. Kolokace jako takové nejsou symetrické a záleží na pořadí jejich komponent, resp. na orientaci syntaktické závislosti.

Z důvodu zavedení C_1 a C_2 je třeba jemně pozměnit i některé následující definice a vztahy:

- $C(w_1, w_2)$ je frekvence bigramu $w_1 w_2$ uvedená v seznamu kolokujících bigramů.

- N je počet všech bigramů, resp. součet frekvencí všech kolokujících bigramů v seznamu.

$$N = \sum_{ij} C(w_i, w_j) = C(*)$$

základní tvar	frekvence
(,, :----)	88 830
(., :----)	81 098
(-, :----)	49 962
(a, ^----)	36 725
(být, B--A-)	32 922
(v, R----)	22 913
(se, 7----)	22 360
(na, R----)	21 394
(\&, :----)	14 534
(), :----)	13 958

a)

základní tvar	frekvence
(být, B--A-)	32 922
(systém, NI-A-)	9 781
(firma, NF-A-)	8 484
(počítač, NI-A-)	7 110
(program, NI-A-)	5 300
(síť, NF-A-)	4 547
(uživatel, NM-A-)	4 125
(být, c----)	3 921
(mít, B--A-)	3 774
(aplikace, NF-A-)	3 409

b)

Tabulka 6.1: Nejpočetnější základní tvary slov v kolekci $\mathcal{CW}94$: a) celkem, b) pouze autosémantické slovní druhy (N, A, C, V, D)

- $C_1(w_1)$ je součet frekvencí všech bigramů, jejichž *první* slovo je w_1.

$$C_1(w_1) = \sum_j C(w_1, w_j) = C(w_1, *)$$

- $C_2(w_2)$ je součet frekvencí všech bigramů, jejichž *druhé* slovo je w_2.

$$C_2(w_2) = \sum_i C(w_i, w_2) = C(*, w_2)$$

- $P_1(w_1)$ je maximální věrohodný odhad pravděpodobnosti výskytu slova w_1 na *prvním* místě bigramu.

$$P_1(w_1) = \frac{C_1(w_1)}{N}$$

- $P_2(w_2)$ je maximální věrohodný odhad pravděpodobnosti výskytu slova w_2 na *druhém* místě bigramu.

$$P_2(w_2) = \frac{C_2(w_2)}{N}$$

- Nulová hypotéza o nezávislosti výskytu slov w_1 a w_2 pro testování hypotéz *t testem* a χ^2 *testem* je definována takto:

$$H_0 : P(w_1 w_2) = P_1(w_1) P_2(w_2)$$

- Pro test poměrem pravděpodobností používáme stejné hypotézy:

$$\begin{aligned} H_1 : &\quad P(w_2|w_1) = p = P(w_2|\neg w_1) \\ H_2 : &\quad P(w_2|w_1) = p_1 \neq p_2 = P(w_2|\neg w_1), \end{aligned}$$

kde platí:

$$P(w_2|w_1) = \frac{P(w_1w_2)}{P_1(w_1)}$$
$$P(w_2|\neg w_1) = \frac{P(\neg w_1w_2)}{P_1(\neg w_1)}$$

Ve vztahu

$$\lambda = \frac{L(H_1)}{L(H_2)} = \frac{b(c_{12};c_1,p)b(c_2-c_{12};N-c_1,p)}{b(c_{12};c_1,p_1)b(c_2-c_{12};N-c_1,p_2)},$$

jsou pak:

$$c_1 = C_1(w_1) \qquad p = \frac{c_2}{N}$$
$$c_2 = C_2(w_2) \qquad p_1 = \frac{c_{12}}{c_1}$$
$$c_{12} = C(w_1w_2) \qquad p_2 = \frac{c_2-c_{12}}{N-c_1}$$

- Vztah pro výpočet vzájemné informace se obmění analogicky:

$$I(x,y) = log_2\frac{P(w_1w_2)}{P_1(w_1)P_2(w_2)}$$

Nyní se budeme detailně zabývat jednotlivými experimenty a rozdíly mezi nimi. Především se jedná o rozdílné definice pojmu *společný výskyt dvou slov* a také o použité metody filtrace bigramů, které nejsou kolokacemi.

6.1.1 Slova sousedící

Společný výskyt sousedících slov v povrchovém slovosledu w_1w_2 nastává, pokud se v textu nachází slovo w_2 těsně za slovem w_1. V případě, že uvažujeme každou dvojici sousedících slov jako kolokující bigram, bude postup výpočtů u všech metod stejný, jako byl popsán v kapitolách 5.1 - 5.4.

Každý konkrétní výskyt slovního tvaru v textu je součástí dvou bigramů, proto jsou hodnoty C_1 a C_2 shodné pro všechna w, až na první a poslední slovo v textu.

$$C_1(w) = C(w,*) = C(*,w) = C_2(w)$$

Absolutní vzdálenosti komponent bigramů jsou vždy rovny 1, střední hodnotu a rozptyl tudíž nemá smyl počítat. Po aplikaci těchto metod použijeme *slovnědruhový filtr* a vzory z tabulky 5.20 rozšířené o slovní poddruhy z tabulky 5.21. Předpokládáme, že touto metodou lze odhalovat především *pevné kolokace*.

6.1.2 Eliminace stopslov

Metoda eliminace stopslov upravuje seznam získaný předchozí metodou. Odstraňuje z něj všechny bigramy, jejichž *obě* komponenty jsou *stopslova*, tedy slova, která nevytvářejí kolokace

$w_1 w_2$	C^{-5}	C^{-4}	C^{-3}	C^{-2}	C^{-1}	C^0	C^1	C^2	C^3	C^4	C^5
jak být	2	2	2	1	6	0	0	0	2	3	0
být jak	0	3	2	0	0	0	6	1	2	2	2

Tabulka 6.2: Frekvence výskytu bigramů pro jednotlivé vzdálenosti. V tabulce je ten samý bigram dvakrát, pouze má obrácené pořadí komponent. Při extrakci kolokujících bigramů pomocí kolokačního okénka je vhodné se vyvarovat dvojímu zpracování jednoho bigramu.

(viz kapitola 4.4). Veškeré výpočty se budou provádět stejně jako v předchozím případě. Střední hodnota a rozptyl vzdáleností komponent kolokací opět zkoumat nebudeme. Pouze počet kolokací N bude nižší. Experimenty ukazují, že toto snížení se obvykle pohybuje na úrovní asi 10 %.

Jedná se pouze o vylepšení předchozího způsobu získávání kolokujících bigramů. Stejně jako v předchozím případě jsme schopni odhalovat spíše *pevné kolokace*. Předpokládáme ale zlepšení výsledků zmenšením počtu zkoumaných kolokací a jejich vyčištěním. Eliminaci stopslov budeme aplikovat i na seznamy kolokujících bigramů získané všemi dalšími způsoby.

Po aplikování všech metod použijeme také *slovnědruhový filtr* a odstraníme všechny dvojice slov, které neodpovídají *slovnědruhovým vzorům*, stejně jako v předchozím případě.

6.1.3 Kolokační okénko

U bigramů, které získáme pomocí kolokačního okénka, můžeme sledovat i jednotlivé počty jejich výskytů pro všechny možné vzdálenosti slov. Velikost kolokačního okénka zvolíme $n = 5$.

Tato metoda je přímo určená pro měření střední hodnoty a rozptylu vzdálenosti slov kolokace. Můžeme však použít i další metody. Společný výskyt je v tomto případě definován jako výskyt v rámci jednoho kolokačního okénka. Počet společných výskytů slov w_1 a w_2 je tedy definován takto:

$$C(w_1 w_2) = \sum_{i=-5}^{5} C^i(w_1 w_2),$$

kde $C^i(w_1 w_2)$ je počet výskytů bigramu $w_1 w_2$ se vzdáleností komponent i.

Frekvence jednotlivých slov $C_1(w)$ a $C_2(w)$ jsou pak definovány s použitím předchozí definice. Opět pro ně platí, že jsou si pro všechny w rovné. Tentokrát je to z důvodu, že u bigramů získaných kolokačním okénkem nerozlišujeme pořadí komponent. Viz ilustrativní příklad v tabulce 6.2.

$$C_1(w) = C(w, *) = C(*, w) = C_2(w)$$

Předpokládaná výhoda metod, které pracují s bigramy získanými pomocí kolokačního okénka je především v tom, že budou odhalovat více1 *volných kolokací*. Filtrovat budeme kolokace nevhodných slovních druhů podobně jako v předchozích případech.

6.1.4 Syntaktická závislost

Pro tuto metodu platí, že dvě slova mají společný výskyt právě tehdy, pokud jsou syntakticky závislá. Znalost syntaktických závislostí ve větě nám umožňuje pracovat pouze s bigramy, o kte-

rých víme, že mezi jejich komponentami existuje syntaktická závislost. Všechny kolokace *1. druhu*, na které je tato metoda zaměřená, tuto závislost nutně obsahují. Je proto velmi výhodné vůbec se nezabývat bigramy, jejichž slova syntakticky asociovaná nejsou.

Seznam kolokujících bigramů získaný jako syntakticky závislé dvojice nám dovoluje použití všech metod detekce kolokací kromě metody střední hodnoty a rozptylu, s použitím vztahů uvedených na začátku kapitoly. Pokud máme možnost pracovat i s typem syntaktické závislosti, pak pro *t test* a χ^2 *test* bude nulová hypotéza znít takto (viz kapitola 5.5):

$$H_0 : P(w_1, rel, w_2) = P(rel)P(w_1|rel)P(w_2|rel)$$

Stejný odhad pravděpodobnosti společného nezávislého výskytu slov w_1 a w_2 s typem závislosti rel použijeme při výpočtu vzájemné informace:

$$I(w_1, rel, w_2) = log_2 \frac{P(w_1, rel, w_2)}{P(rel) \times P(w_1|rel) \times P(w_2|rel)}$$

Pro odstranění bigramů, o kterých víme, že netvoří kolokace, použijeme kromě slovnědruhových vzorů z tabulky 5.20 a 5.21 i typy syntaktických závislosti z tabulky 5.22.

6.1.5 Tranzitivní syntaktická závislost

Tranzitivní syntaktickou závislost definovanou v kapitole 4.3 použijeme pro sestavování seznamu kolokujících bigramů, které budou kandidovat především na kolokace *2. druhu*, tedy slova sémanticky podobná a příbuzná (viz kapitola 4.4). Společným výskytem dvou slov rozumíme situaci, kdy jsou tato slova v relaci tranzitivní syntaktické závislosti.

Tento typ nově definované závislosti slov umožňuje sledovat i tzv. *vzdálenost v závislostním podstromu* (viz kapitola 4.4). Pro všechny tranzitivně syntakticky závislé dvojice slov můžeme počítat i jejich vzdálenost ve výše uvedeném smyslu. Seznam kolokujících bigramů pak bude obsahovat frekvence bigramů pro jednotlivé vzdálenosti jejich komponent. Parametr této metody n, tedy maximální vzdálenost v syntaktickém stromu, s kterou budeme pracovat, volíme $n = 4$.

Tento způsob získávání kolokujících bigramů poskytuje data i pro aplikaci metody detekci kolokací pomocí měření odhadů střední hodnoty a rozptylu takto definované vzdálenosti slov. Evidentně se v tomto postupu projevuje souvislost s metodou *kolokačního okénka*, která používala povrchovou vzdálenost slov ve větě.

V ostatních metodách detekce kolokací v posloupnosti bigramů s tranzitivní syntaktickou závislostí použijeme opět definici $C(w_1w_2)$:

$$C(w_1w_2) = \sum_{i=-4}^{4} C^i(w_1w_2),$$

kde $C^i(w_1w_2)$ je počet výskytů bigramu w_1w_2 se vzdálenosti jeho komponent rovné i.

Filtrovat budeme pouze podle slovních druhů a poddruhů z tabulek 5.20 a 5.21.

6.1.6 Sourozenecký syntaktický vztah

Relace sourozeneckého syntaktického vztahu definovaná v kapitole 4.4 slouží k získání bigramů, které jsou kandidáty především na kolokace *2. druhu*. Jejich komponenty nejsou syntakticky asociované, ale rozvíjejí jedno slovo, což je znak charakteristický pro kolokací *2. druhu*, a tím spíše pokud je typ jejich syntaktické závislosti na tomto slově stejný (viz kapitola 5.6). Ostatní bigramy pak odstraníme ze seznamu kandidátů na kolokace. Seznam kolokujících bigramů získaných touto metodou použijeme pro všechny metody detekce kolokací kromě střední hodnoty a rozptylu.

6.2 Nástroje

Při realizaci všech experimentů jsme brali na zřetel především následující skutečnosti či požadavky a podle nich byly zvoleny i použité nástroje.

- formát zpracovávaných textů – SGML
- jednoduchý a přehledný zápis použitých algoritmů
- velmi velké množství zpracovávaných a výsledných dat
- možnost jednoduchého a přehledného přístupu k datům
- nezávislost na použité výpočetní platformě

Neklademe žádné požadavky na optimalizaci, protože naším cílem není implementace metody s minimální paměťovou a výpočetní složitostí. Jde nám o výsledky, nikoliv o náročnost jejich získání, přestože k jejímu snížení některé postupy vedou (např. eliminace stopslov a filtrace bigramů).

6.2.1 Výpočty

Jako nejvýhodnější výpočetní prostředek vyhovující našim požadavkům jsme určili programovací jazyk *Perl*. Ten umožňuje velmi snadnou práci s texty, snadný zápis algoritmů a matematických výpočtů a prostřednictvím modulů jednoduchý přístup k datovým uložištím. Jednotlivé metody jsou implementovány ve více skriptech. Práce s těmito skripty, které na sebe různě navazují, je zjednodušena použitím utility `make`. Vhodnou volbou jejích parametrů lze řídit průběh všech experimentů.

6.2.2 Data

Pro zjednodušení práce s daty jsme jako jejich úložiště v průběhu výpočtů a analýzy zvolili databázi *PostgreSQL*. V kombinací s jazykem *Perl* tvoří silný nástroj pro studium metod NLP.

Do databáze ukládáme veškerá získaná a spočtená data včetně různých statistik:

- statistické informace o jednotlivých dokumentech
- slovní tvary

- základní slovní tvary
- kolokující bigramy šesti různých metod
- statistické informace o seznamech kolokujících bigramů
- výsledky šesti metod detekce kolokací pro šest různých způsobů extrakce bigramů

Pro lepší analýzu výsledků dat byla vytvořena aplikace sloužící k prohlížení obsahu použité databáze. Za účelem rychlého vývoje a minimalizace práce s grafickým rozhraním byla zvolena platforma WWW aplikace. Na straně klienta libovolný WWW prohlížeč, na straně serveru WWW server (*Apache*) s modulem pro jazyk *PHP* s podporou přístupu k databázi *PostgreSQL*. Vše implementováno v jazyce *PHP* s použitím knihovny *Fast Template*.

Uživatel má možnost velmi komfortním způsobem prohlížet data uvedená v předchozím seznamu, třídit je podle různých atributů, filtrovat podle jejich hodnot atd. Objem uložených dat je v případě rozsáhlých kolekcí dokumentů velmi velký, spokojenost uživatele tak závisí především na rychlosti serverové strany aplikace a rychlosti spojení. Na obrázku 6.2 je obrazovka WWW prohlížeče sejmutá ve chvíli, kdy jsou zobrazovány výsledky metod detekce kolokací aplikovaných na seznam bigramů získaných jako závislostní dvojice. Seznam bigramů je setříděn sestupně podle spočtené hodnoty statistiky χ^2 a obsahuje pouze kolokace slovnědruhového typu AN, s počtem výskytů minimálně 3 a frekvencí závislého slova větší než 5.

6.2.3 Ohodnocování detekovaných kolokací

Jelikož pro analýzu výsledků různých metod detekce kolokací chceme použití hodnocení kolokací lidskými silami, bylo nutné proces toho hodnocení nějakým způsobem podpořit a zjednodušit.

Nejdříve byla vytvořena GUI aplikace v jazyce *C++* s použitím knihovny *Qt*, která pracuje se seznamem bigramů získaných z databáze pomocí příkazu `make exportbigrams`, zobrazuje jednotlivé bigramy a od uživatele čeká rozhodnutí o zařazení do odpovídající kategorie provedené stisknutím příslušného tlačítka. Ten také vidí, jaký podíl bigramů je již ohodnocených a kolik jich ještě zbývá. V případě chybného určení se může vrátit až o 5 kroků zpět a hodnocení změnit. Další funkcí programu je možnost změnit základní tvar kolokace, pokud není správně určen. Každý bigram je totiž reprezentován dvojicí základních tvarů komponent, typem syntaktické závislosti (v případě, že je znám) a řetězcem, který tvoří základní tvar celého bigramu. Určení tohoto slovníkového tvaru kolokace je součástí procedur jednotlivých metod. Okno této miniaplikace je na obrázku 6.3.

Práce s touto aplikací se však ukázala neefektivní. Žádným způsobem totiž nevyužívala výsledků získaných v provedených experimentech. Nabídlo se řešení, jak tyto výsledky použít a proces hodnocení kolokací člověku ulehčit.

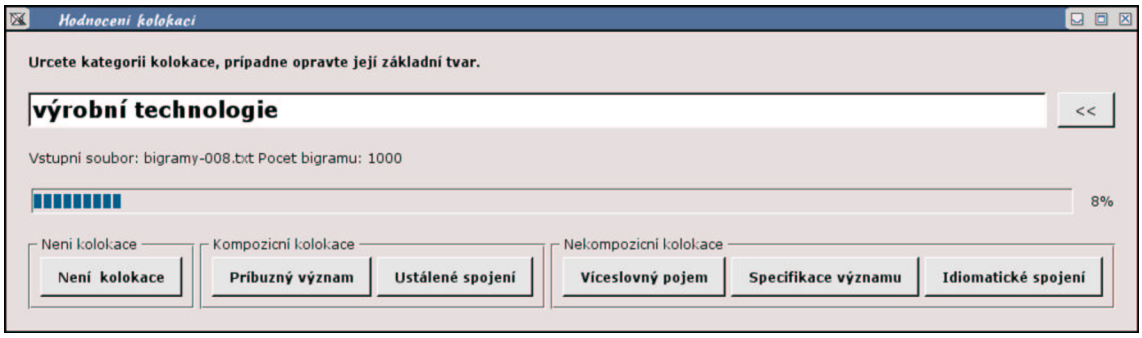
Aplikace původně určená pouze k prohlížení dat uložených v databázi byla rozšířena právě o možnost zařazování kolokací do příslušných kategorií. Obrovská výhoda spočívá v tom, že si uživatel může jakýmkoliv způsobem setřídit a filtrovat bigramy podle výsledků jednotlivých metod a tím proces hodnocení zjednodušit. Pokud bude například pracovat se seznamem bigramů slovnědruhového typu AN (tedy nejčastější typ kolokací), které neprošly *t testem* (nulová hypotéza zamítnuta), bude jeho práce spočívat pouze v kontrole, zda některý bigram kolokací není. Obrazovka prohlížeče, který očekává hodnocení kolokací je na obrázku 6.4.

Seznam všech bigramů získaných metodou "d"

1 ok reset Počet vyhovujících: 50 Počet stran: 1 Aktuální strana: 1

Kolokace	SC	Lemma 1	Po	Ge	Gr	Ng	Va	Freq	Prob	Lemma 2	Po	Ge	Gr	Ng	Va	Freq	Prob	PC	Freq	t	x2	Ir	11	StaC	ColC
																>5		NA	>3	>2.7	>3.8			0	
ručení omezený	N:Atr:A	ručení	N	N	-	A	8	11	0.00032	omezený	A	N	1	A	8	11	0.00032	NA	11	0.00084	55.947	3152.000	256.424	8.163	0	
papír cenný	N:Atr:A	papír	N	I	-	A	-	9	0.00026	cenný	A	I	1	A	-	7	0.00020	NA	7	0.00054	56.018	3152.000	158.555	8.815	0	
znalec soudní	N:Atr:A	znalec	N	M	-	A	-	12	0.00034	soudní	A	M	1	A	-	6	0.00017	NA	6	0.00046	51.863	2700.856	126.365	8.815	0	
poradce daňový	N:Atr:A	poradce	N	M	-	A	-	10	0.00029	daňový	A	M	1	A	-	6	0.00017	NA	6	0.00046	51.863	2700.856	130.946	8.815	0	
zůstatek likvidační	N:Atr:A	zůstatek	N	I	-	A	-	8	0.00023	likvidační	A	I	1	A	-	7	0.00020	NA	7	0.00054	49.372	2449.997	163.614	8.452	0	
unie celní	N:Atr:A	unie	N	F	-	A	-	8	0.00023	celní	A	F	1	A	-	9	0.00026	NA	7	0.00054	48.982	2411.496	149.859	8.429	0	
galerie národní	N:Atr:A	galerie	N	F	-	A	-	11	0.00032	národní	A	F	1	A	-	7	0.00020	NA	6	0.00046	47.998	2314.037	120.214	8.592	0	
obrat roční	N:Atr:A	obrat	N	I	-	A	-	7	0.00020	roční	A	I	1	A	-	7	0.00020	NA	5	0.00038	47.344	2249.998	103.768	8.815	0	
voda teplý	N:Atr:A	voda	N	F	-	A	-	12	0.00034	teplý	A	F	1	A	-	11	0.00032	NA	10	0.00077	45.622	2097.989	210.595	7.715	0	
daň silniční	N:Atr:A	daň	N	F	-	A	-	23	0.00066	silniční	A	F	1	A	-	12	0.00034	NA	10	0.00077	44.728	2016.525	174.835	7.659	0	
privatizace kuponový	N:Atr:A	privatizace	N	F	-	A	-	11	0.00032	kuponový	A	F	1	A	-	6	0.00017	NA	5	0.00038	43.200	1873.925	98.267	8.552	0	
strana druhý	N:Atr:A	strana	N	F	-	A	-	10	0.00029	druhý	A	F	1	A	-	18	0.00052	NA	9	0.00069	41.893	1769.051	169.637	7.622	0	
ruch cestovní	N:Atr:A	ruch	N	I	-	A	-	7	0.00020	cestovní	A	I	1	A	-	6	0.00017	NA	5	0.00038	40.394	1638.954	108.051	8.359	0	
pacient tuzemský	N:Atr:A	pacient	N	M	-	A	-	15	0.00043	tuzemský	A	M	1	A	-	6	0.00017	NA	4	0.00031	34.537	1197.710	68.810	8.230	0	
bilance platební	N:Atr:A	bilance	N	F	-	A	-	4	0.00011	platební	A	F	1	A	-	6	0.00017	NA	4	0.00031	34.537	1197.710	93.904	8.230	0	
rozpočet státní	N:Atr:A	rozpočet	N	I	-	A	-	8	0.00023	státní	A	I	1	A	-	10	0.00029	NA	5	0.00038	32.931	1090.340	92.667	7.774	0	
republika česky	N:Atr:A	republika	N	F	-	A	-	23	0.00066	český	A	F	1	A	-	50	0.00144	NA	18	0.00138	31.724	1028.821	265.031	5.855	0	
partner obchodní	N:Atr:A	partner	N	M	-	A	-	18	0.00052	obchodní	A	M	1	A	-	8	0.00023	NA	6	0.00046	31.641	1008.204	104.331	7.400	0	
tah hlavní	N:Atr:A	tah	N	I	-	A	-	8	0.00023	hlavní	A	I	1	A	-	12	0.00034	NA	5	0.00038	30.465	933.743	89.152	7.552	0	
produkt domácí	N:Atr:A	produkt	N	I	-	A	-	14	0.00040	domácí	A	I	1	A	-	6	0.00017	NA	5	0.00038	29.421	871.382	93.808	7.452	0	
veletrh mezinárodní	N:Atr:A	veletrh	N	I	-	A	-	14	0.00040	mezinárodní	A	I	1	A	-	8	0.00023	NA	5	0.00038	29.257	861.421	86.342	7.436	0	
obrázek malý	N:Atr:A	obrázek	N	I	-	A	-	21	0.00060	malý	A	I	2	A	-	6	0.00017	NA	4	0.00031	28.854	836.795	64.408	7.715	0	
růst ekonomický	N:Atr:A	růst	N	I	-	A	-	38	0.00109	ekonomický	A	I	1	A	-	14	0.00040	NA	8	0.00061	28.011	792.615	109.868	6.645	0	
vlna druhý	N:Atr:A	vlna	N	F	-	A	-	11	0.00032	druhý	A	F	1	A	-	18	0.00052	NA	4	0.00031	27.929	784.997	56.629	7.622	0	
trh kapitálový	N:Atr:A	trh	N	I	-	A	-	52	0.00149	kapitálový	A	I	1	A	-	6	0.00017	NA	5	0.00038	27.893	784.244	72.650	7.300	0	
výstava mezinárodní	N:Atr:A	výstava	N	F	-	A	-	19	0.00055	mezinárodní	A	F	1	A	-	10	0.00029	NA	5	0.00038	27.893	782.992	76.354	7.300	0	
výkon skutečný	N:Atr:A	výkon	N	I	-	A	-	14	0.00040	skutečný	A	I	1	A	-	6	0.00017	NA	4	0.00031	27.498	760.237	69.741	7.578	0	
centrum manažerský	N:Atr:A	centrum	N	N	-	A	-	8	0.00023	manažerský	A	N	1	A	-	6	0.00017	NA	4	0.00031	27.498	760.237	77.906	7.578	0	
stanice telefonní	N:Atr:A	stanice	N	F	-	A	-	11	0.00032	telefonní	A	F	1	A	-	7	0.00020	NA	5	0.00038	27.212	745.706	93.985	7.230	0	
fond manažerský	N:Atr:A	fond	N	I	-	A	-	28	0.00080	manažerský	A	I	1	A	-	6	0.00017	NA	6	0.00046	25.758	670.705	108.387	6.815	0	
fond svazový	N:Atr:A	fond	N	I	-	A	-	28	0.00080	svazový	A	I	1	A	-	6	0.00017	NA	6	0.00046	25.758	670.705	108.387	6.815	0	
země evropský	N:Atr:A	země	N	F	-	A	-	23	0.00066	evropský	A	F	1	A	-	10	0.00029	NA	5	0.00038	25.430	651.242	73.207	7.037	0	
síť telekomunikační	N:Atr:A	síť	N	F	-	A	-	16	0.00046	telekomunikační	A	F	1	A	-	6	0.00017	NA	4	0.00031	24.339	596.185	67.948	7.230	0	
cena regulovaný	N:Atr:A	cena	N	F	-	A	-	124	0.00356	regulovaný	A	F	1	A	-	7	0.00020	NA	7	0.00054	20.890	444.272	94.639	6.007	0	
místo pracovní	N:Atr:A	místo	N	N	-	A	-	28	0.00080	pracovní	A	N	1	A	-	6	0.00017	NA	4	0.00031	20.305	415.729	60.781	6.715	0	
rok loňský	N:Atr:A	rok	N	I	-	A	-	126	0.00362	loňský	A	I	1	A	-	9	0.00026	NA	6	0.00046	19.547	387.116	64.314	6.037	0	
doba pracovní	N:Atr:A	doba	N	F	-	A	-	50	0.00144	pracovní	A	F	1	A	-	14	0.00040	NA	7	0.00054	19.497	385.242	85.954	5.815	0	
komora hospodářský	N:Atr:A	komora	N	F	-	A	-	19	0.00055	hospodářský	A	F	1	A	-	12	0.00034	NA	4	0.00031	18.567	347.357	54.697	6.462	0	
rok letošní	N:Atr:A	rok	N	I	-	A	-	126	0.00362	letošní	A	I	1	A	-	7	0.00020	NA	5	0.00038	18.489	346.132	55.214	6.137	0	
ekonomika český	N:Atr:A	ekonomika	N	F	-	A	-	32	0.00092	český	A	F	1	A	-	50	0.00144	NA	11	0.00084	17.798	324.138	121.174	4.943	0	
společnost veřejný	N:Atr:A	společnost	N	F	-	A	-	60	0.00172	veřejný	A	F	1	A	-	6	0.00017	NA	5	0.00038	17.844	323.117	70.489	6.037	0	
doba poslední	N:Atr:A	doba	N	F	-	A	-	50	0.00144	poslední	A	F	1	A	-	6	0.00017	NA	4	0.00031	17.095	295.425	53.707	6.230	0	
firma soukromý	N:Atr:A	firma	N	F	-	A	-	119	0.00342	soukromý	A	F	1	A	-	7	0.00020	NA	5	0.00038	14.374	210.532	56.053	5.436	0	
společnost leasingový	N:Atr:A	společnost	N	F	-	A	-	60	0.00172	leasingový	A	F	1	A	-	6	0.00017	NA	4	0.00031	14.220	205.199	51.526	5.715	0	
firma zahraniční	N:Atr:A	firma	N	F	-	A	-	119	0.00342	zahraniční	A	F	1	A	-	9	0.00026	NA	5	0.00038	12.591	161.645	50.348	5.074	0	
společnost obchodní	N:Atr:A	společnost	N	F	-	A	-	60	0.00172	obchodní	A	F	1	A	-	18	0.00052	NA	6	0.00046	12.453	157.932	61.170	4.798	0	
společnost lázeňský	N:Atr:A	společnost	N	F	-	A	-	60	0.00172	lázeňský	A	F	1	A	-	9	0.00026	NA	4	0.00031	11.498	134.286	44.743	5.130	0	
informace další	N:Atr:A	informace	N	F	-	A	-	41	0.00118	další	A	F	1	A	-	28	0.00080	NA	4	0.00031	10.526	112.290	34.167	4.891	0	
firma malý	N:Atr:A	firma	N	F	-	A	-	119	0.00342	malý	A	F	1	A	-	13	0.00037	NA	4	0.00031	8.964	81.987	31.497	4.462	0	
firma český	N:Atr:A	firma	N	F	-	A	-	119	0.00342	český	A	F	1	A	-	50	0.00144	NA	6	0.00046	5.774	34.437	29.947	2.892	0	

Obrázek 6.2: WWW prohlížeč zobrazující výsledky metod aplikovaných na syntakticky závislé dvojice slov



Obrázek 6.3: Obrazovka aplikace pro hodnocení kolokací

Hodnocení bigramů získaných metodou "d"

1 ok reset Počet vyhovujících: **50** Počet stran: **1** Aktuální strana: **1**

Kolokace	SC	No	Kompoz	Nekompoz	X	PC	Freq	Prob	t	x2	lr	I1
	N:Atr	0	1 2	3 4 5			>4		>2.78			>5.4
ručení omezený	N:Atr:A					NA	11	0.00084	55.947	3152.000	256.424	8.163
papír cenný	N:Atr:A					NA	7	0.00054	56.018	3152.000	158.555	8.815
znalec soudní	N:Atr:A					NA	6	0.00046	51.863	2700.856	126.365	8.815
poradce daňový	N:Atr:A					NA	6	0.00046	51.863	2700.856	130.946	8.815
teplárna moravskoslezský	N:Atr:A					NA	5	0.00038	51.153	2625.832	115.851	9.037
družstvo bytový	N:Atr:A					NA	5	0.00038	51.153	2625.832	94.677	9.037
zůstatek likvidační	N:Atr:A					NA	7	0.00054	49.372	2449.997	163.614	8.452
unie celní	N:Atr:A					NA	7	0.00054	48.982	2411.496	149.859	8.429
galerie národní	N:Atr:A					NA	6	0.00046	47.998	2314.037	120.214	8.592
bod procentní	N:Atr:A					NA	5	0.00038	47.344	2249.998	104.418	8.815
obrat roční	N:Atr:A					NA	5	0.00038	47.344	2249.998	103.768	8.815
voda teplý	N:Atr:A					NA	10	0.00077	45.622	2097.989	210.595	7.715
daň silniční	N:Atr:A					NA	10	0.00077	44.728	2016.525	174.835	7.659
Složil Jaromír	N:Atr:N					NN	5	0.00038	43.253	1879.998	120.134	8.555
privatizace kuponový	N:Atr:A					NA	5	0.00038	43.200	1873.925	98.267	8.552
společnost ručení	N:Atr:N					NN	11	0.00084	41.850	1770.988	212.424	7.333
strana druhý	N:Atr:A					NA	9	0.00069	41.893	1769.051	169.637	7.622
profit číslo	N:Atr:N					NN	10	0.00077	41.109	1707.261	164.675	7.418
ruch cestovní	N:Atr:A					NA	5	0.00038	40.394	1638.954	108.051	8.359
vlna privatizace	N:Atr:N					NN	5	0.00038	40.027	1610.711	84.205	8.333
lázně Teplice	N:Atr:N					NN	5	0.00038	34.140	1172.806	74.721	7.877
rozpočet státní	N:Atr:A					NA	5	0.00038	32.931	1090.340	92.667	7.774
produkt hrubý	N:Atr:A					NA	5	0.00038	32.260	1047.328	101.607	7.715
republika český	N:Atr:A					NA	18	0.00138	31.724	1028.821	265.031	5.855
ÚNMS SR	N:Atr:N					NN	5	0.00038	31.874	1023.176	92.497	7.681
partner obchodní	N:Atr:A					NA	6	0.00046	31.641	1008.204	104.331	7.400
tah hlavní	N:Atr:A					NA	5	0.00038	30.465	933.743	89.152	7.552
produkt domácí	N:Atr:A					NA	5	0.00038	29.421	871.382	93.808	7.452
veletrh mezinárodní	N:Atr:A					NA	5	0.00038	29.257	861.421	86.342	7.436
Dyba ministr	N:Atr:N					NN	5	0.00038	28.215	802.133	112.664	7.333
růst ekonomický	N:Atr:A					NA	8	0.00061	28.011	792.615	109.868	6.645
miliarda koruna	N:Atr:N					NN	13	0.00100	27.849	791.255	167.542	5.946
trh kapitálový	N:Atr:A					NA	5	0.00038	27.893	784.244	72.650	7.300
výstava mezinárodní	N:Atr:A					NA	5	0.00038	27.893	782.992	76.354	7.300
stanice telefonní	N:Atr:A					NA	5	0.00038	27.212	745.706	93.985	7.230
podnikatel začínající	N:Atr:A					NG	5	0.00038	26.578	712.494	77.306	7.163
fond manažerský	N:Atr:A					NA	6	0.00046	25.758	670.705	108.387	6.815
fond svazový	N:Atr:A					NA	6	0.00046	25.758	670.705	108.387	6.815
země evropský	N:Atr:A					NA	5	0.00038	25.430	651.242	73.207	7.037
podnikatel vynikající	N:Atr:A					NG	5	0.00038	24.228	592.265	69.519	6.900
Dyba Karel	N:Atr:N					NN	5	0.00038	23.543	559.239	107.935	6.818
milión dolar	N:Atr:N					NN	8	0.00061	23.049	539.366	123.488	6.096
ministerstvo hospodářství	N:Atr:N					NN	6	0.00046	21.921	486.093	93.666	6.359
cena regulovaný	N:Atr:A					NA	7	0.00054	20.890	444.272	94.639	6.007
zásobování Ostrava	N:Atr:N					NN	6	0.00046	19.961	403.622	109.098	6.096
společnost veřejný	N:Atr:A					NA	5	0.00038	17.844	323.117	70.489	6.037
firma soukromý	N:Atr:A					NA	5	0.00038	14.374	210.532	56.053	5.436

Ulož hodnocení: ok

Označené bigramy patří do kategorie: Není kolokace (0)

Neoznačené bigramy patří do kategorie: Není kolokace (0)

Obrázek 6.4: Rozšíření WWW aplikace pro prohlížení dat o možnost hodnocení kolokací

Kapitola 7

Analýza výsledků metod detekce kolokací

Zatímco v předchozí kapitole jsme popsali experimenty a nástroje použité k jejich provedení, nyní se budeme zabývat analýzou jejich výsledků. Zpočátku probereme obecněji problém hodnocení výsledků metod v této oblasti, poté prezentujeme výsledky realizovaných experimentů a nakonec se také pokusíme o jejich syntézu.

7.1 Způsoby hodnocení výsledků

Jakékoliv precizní hodnocení různých metod či jejich výsledků je v oblasti, ve které se pohybujeme, komplikované. Přirozený jazyk je velmi složitý, chtěli bychom mu „rozumět" na výpočetní úrovni. Nechceme pracovat se slovy a větami, ale s jejich významy a obsahy (například v DIS). Ale právě to je cílem různých oborů NLP a důvod, proč se zabýváme problémy jako je detekce kolokací a mnoho dalších.

Existují dva naprosto rozdílné přístupy, jak posuzovat kvalitu řešení naší úlohy: *subjektivní hodnocení* a *hodnocení dle výsledků cílové aplikace*.

7.1.1 Hodnocení dle výsledků cílové aplikace

Předpokládejme, že jsme problém detekce kolokací řešili s jedním konkrétním úmyslem, pro jednu tzv. *cílovou aplikaci*. Potom můžeme kvalitu našeho řešení hodnotit podle výsledků této cílové aplikace. Předpokládáme, že tyto výsledky hodnotit umíme. Jde tedy o jakési převedení jednoho problému na druhý.

Příklad takové cílové aplikace může být dokumentografický informační systém (viz kapitola 8), který pracuje s jednoslovnými termy. Jeho kvalitu popisují koeficienty *úplnosti* a *přesnosti)*:

Koeficient úplnosti (R, Recall):

$$R = \frac{N_{vr}}{N_r} \in \langle 0, 1 \rangle$$

Koeficient přesnosti (P, Precision):

$$P = \frac{N_{vr}}{N_v} \in \langle 0, 1 \rangle$$

Kde

N_{vr} = počet vyhledaných relevantních dokumentů,
N_r = počet relevantních dokumentů,
N_v = počet vyhledaných dokumentů.

Tedy R reprezentuje pravděpodobnost, že relevantní dokument bude vybrán, zatímco P představuje pravděpodobnost, že vyhledaný dokument je relevantní. V ideálním případě jsou oba koeficienty rovny 1.

Určení *koeficientu přesnosti* tedy spočívá ve zkoumání všech dokumentů, které systém vyhledal a zhodnocení, zda odpovídají požadavkům uživatele (relevantnost). K přesnému určení *koeficientu úplnosti* je potřeba analyzovat všechny dokumenty v systému, což je mnohem náročnější.

Úspěšnost metod detekce kolokací můžeme tedy měřit podle rozdílů hodnot R a P systémů vyhledávání informací, které s kolokacemi nepracují a těch, které je naopak využívají. Takové systémy jsme však neměli k dispozici, a proto budeme muset naše výsledky hodnotit jinak.

7.1.2 Subjektivní hodnocení

Subjektivní hodnocení je oproti předchozímu případu daleko prozaičtější. Úlohu vyhledávání kolokací v textech vyřešíme lidskou silou a takto získané výsledky porovnáme s výsledky získanými metodami automatickými. K formálnímu měření úspěšnosti hledání kolokací v textech použijeme také koeficienty *úplnosti* a *přesnosti* definované takto:

Koeficient úplnosti (R, Recall):

$$R = \frac{N_{vk}}{N_k} \in \langle 0, 1 \rangle$$

Koeficient přesnosti (P, Precision):

$$P = \frac{N_{vk}}{N_v} \in \langle 0, 1 \rangle$$

Kde

N_{vk} = počet detekovaných bigramů, které jsou kolokacemi,
N_k = počet všech kolokací,
N_v = počet detekovaných bigramů.

Tedy R představuje pravděpodobnost, že kolokace bude odhalena, zatímco P představuje pravděpodobnost, že detekované bigramy jsou kolokace. Ideálem je, stejně jako u dokumentografických systémů, aby oba koeficienty byly rovny 1.

7.2 Prakticky dosažené výsledky

Jelikož jsme neměli k dispozici žádnou cílovou aplikaci, pokusili jsme se srovnat výsledky metod detekce kolokací podle různých způsobů získávání kolokujících bigramů subjektivní metodou.

7.2.1 Získání ohodnocených kolokací

Nemalým problémem byl výběr bigramů, které měly být ohodnoceny. Vybraný seznam bigramů měl být dostatečně reprezentativní vzhledem k celku, aby výsledky nebyly zkreslené, dostatečně malý, aby celý proces nebyl příliš časově náročný, a také neměl obsahovat bigramy se zkreslenými hodnotami provedených výpočtů. Vzorek jsme sestavili takto: Všech šest různých způsobů získání kolokujících bigramů jsme aplikovali na kolekci $\mathcal{CW}94$ a ve všech šesti případech jsme vybrali všechny dvojice s frekvencí výskytu vetší než 5. U metod, které používají kolokační okénko, jsme požadovali, aby frekvence překročila tuto hodnotu alespoň pro jednu z možných vzdáleností komponent. Získané množiny dvojic jsme sjednotili a takto získané dvojice slov jsme pomocí aplikace popsané v kapitole 6.2 ohodnotili. Tímto způsobem bylo zpracováno přesně 12780 bigramů, z nichž 41 % bylo ohodnoceno jako kolokace. Zastoupení jednotlivých druhů kolokací viz tabulka 7.1.

typ kolokace	**zastoupení**
kolokace kompoziční	
1. sémanticky příbuzná slova	4,7 %
2. ustálená slovní spojení	16,9 %
kolokace nekompoziční	
3. slabě nekompoziční	72,4 %
4. částečně nekompoziční	5,8 %
5. dokonale nekompoziční	0,2 %

Tabulka 7.1: Poměr zastoupení různých druhů kolokací detekovaných v kolekci $\mathcal{CW}94$

Zdůrazněme, že ohodnocování kolokací a jejich klasifikace do různých kategorií je velice subjektivní a obtížné. Nelze říct, do jaké míry je naše ohodnocení přesné. Důležité však je, že pro hodnocení úspěšnosti různých metod detekce kolokací byla použita stejná množina lidskou silou ohodnocených kolokací.

Na následujících stranách srovnáme šest různých variant extrakce kolokujících bigramů. Budeme je značit následujícím způsobem:

1. slova sousedící v povrchovém slovosledu
2. slova sousedící v povrchovém slovosledu s eliminací stopslov
3. kolokační okénko v povrchovém slovosledu velikosti $n = 5$
4. slova syntakticky závislá

	1	**2**	**3**	**4**	**5**	**6**
kolokující bigramy	621 194	483 726	2 169 441	525 614	1 391 949	435 487
bez chybně určených bigramů	575 589	448 717	1 986 581	468 077	1 218 678	410 237
po filtraci nevhodných sl. dr.	286 168	280 924	1 336 592	276 596	785 547	216 681
po filtraci slovnědruh. vzorů	253 524	248 499	1 195 201	251 537	667 309	195 477
po aplikaci testů	[a)]220683	[b)]213 858	[c)]1 057 212	211 065	[d)]616 706	[e)]155 350

kolokující bigramy	100 %	100 %	100 %	100 %	100 %	100 %
bez chybně určených bigramů	92 %	92 %	91 %	89 %	87 %	94 %
po filtraci nevhodných sl. dr.	46 %	58 %	61 %	52 %	56 %	49 %
po filtraci slovnědruh. vzorů	40 %	51 %	55 %	47 %	47 %	44 %
po aplikaci testů	35 %	44 %	48 %	40 %	44 %	33 %

Tabulka 7.2: Absolutní (první tabulka) a relativní (druhá tabulka - v procentech) výsledky metod detekce kolokací aplikovaných na 6 způsobů získávání kolokujících bigramů

5. slova tranzitivně syntakticky závislá do vzdálenosti $n = 4$
6. slova v sourozeneckém syntaktickém vztahu

7.2.2 Výsledky dle variant získání kolokujících bigramů

V tabulce 7.2 jsou uvedeny souhrnné výsledky metod 1 - 6 aplikovaných na kolekci $\mathcal{CW}94$, tak jak jsme je popsali v kapitole 6. V první části jsou absolutní počty bigramů, v druhé části jsou pak ty samé výsledky v procentech relativně ke všem nalezeným kolokujícím bigramům.

Pro každou z metod obsahují následující údaje: V prvním řádku jsou počty všech extrahovaných bigramů, v druhém řádku pak počty kolokujících bigramů po odstranění chyb způsobených parserem. Na dalších dvou řádcích jsou údaje o počtech bigramů, které zůstaly v seznamu kandidátů na kolokace po odfiltrování bigramů obsahujících slova nevhodných slovních druhů a bigramů s nevhodnou kombinací slovních druhů komponent. V posledním kroku byly z tohoto seznamu odstraněny bigramy, které byly všemi testy označeny jako nevhodní kandidáti na kolokace. Počet prvků těchto seznamů je na posledním řádku.

Varianty 1 a 2 jsou velmi podobné. V prvním případě se počítají všechna sousedící slova v povrchovém slovosledu. V druhém případě jsou zanedbávány dvojice, jejichž obě komponenty jsou stopslova. Podařilo se tím snížit počet bigramů, které prošly všemi testy a nejsou kolokacemi (viz [a)] a [b)]). Všimněme si zvýšeného počtu kolokujících bigramů u metod 3 a 2 způsobeným použitím kolokačního okénka v povrchovém slovosledu a relací tranzitivní syntaktické závislosti. Objevil se zde velký počet bigramů, mezi jejichž složkami není žádný vztah, který by nasvědčoval, že jde o kolokaci (viz [c)] a [d)]). Nejmenší počet kolokací byl odhalen v seznamu dvojic v sourozeneckém syntaktickém vztahu (viz [e)]), ten by se ještě zmenšil pokud bychom měli v kolekci $\mathcal{CW}94$ k dispozici i typy syntaktické závislosti a mohli tak použít filtraci podle těchto typů.

	1	2	3	4	5	6
bigramy s frekvencí větší než 5	8 060	8 015	12 018	8 288	10 669	2 223
po aplikaci testů	7 456	7 413	9 726	8 009	10 644	1 530
z toho kolokace	3 857	3 845	4 485	4 497	4 601	378
koeficient přesnosti	0,52	0,52	0,46	**0,56**	0,43	0,24
koeficient úplnosti	0,73	0,73	0,85	0,85	**0,88**	0,07

Tabulka 7.3: Absolutní výsledky metod detekce kolokací aplikovaných na 6 způsobů získávání kolokujících bigramů pro bigramy s frekvencí větší než 5

	1	2	3	4	5	6
všech kolokací	100,00 %	100,00 %	100,00 %	100,00 %	100,00 %	100,00 %
kolokace kompoziční						
1. sémanticky příbuzná slova	0,91 %	0,91 %	4,62 %	1,27 %	2,52 %	42,59 %
2. ustálená slovní spojení	17,14 %	17,06 %	15,63 %	16,77 %	16,78 %	5,56 %
kolokace nekompoziční						
3. slabě nekompoziční	74,54 %	74,56 %	73,20 %	76,65 %	75,40 %	35,71 %
4. částečně nekompoziční	7,21 %	7,23 %	6,35 %	5,14 %	5,13 %	15,87 %
5. dokonale nekompoziční	0,21 %	0,23 %	0,20 %	0,18 %	0,17 %	0,26 %

Tabulka 7.4: Zastoupení jednotlivých druhů kolokací (v procentech) ve výsledcích metod detekce kolokací aplikovaných na 6 způsobů získávání kolokačních bigramů pro bigramy s frekvencí větší než 5

V tabulce 7.3 nyní můžeme srovnat jednotlivé varianty extrakce kolokujících bigramů pomocí koeficientů *úplnosti* a *přesnosti*, které jsme spočetli pro bigramy s frekvencí větší než 5. Nejvyššího koeficientu úplnosti dosahuje metoda 5, která získává kolokující bigramy pomocí tranzitivní syntaktické závislosti, a to 88 % (Pokud bychom uvažovali jen kolokace prvého druhu, blížil by se 95 %). Nejvyššího koeficientu přesnosti dosahuje metoda 4, která pracuje se syntakticky závislými dvojicemi. Metoda 6 má oba koeficienty velmi nízké, což je způsobeno jejím zaměřením na *kolokace druhého druhu*.

Na závěr dodejme, že se splnila naše očekávání a jako nejvhodnější se jeví použití variant, které pracují se syntaktickou závislostí. Pokud se zaměříme na *kolokace druhého druhu* pak je nejvhodnější použít variantu 6 a také variantu 2. Celkové srovnání metod podle zastoupení jednotlivých kolokací je obsaženo v tabulkách 7.3 a 7.5.

7.3 Syntéza parciálních výsledků

V předchozí kapitole byly představeny výsledky experimentů popsaných v kapitole 6. Bylo použito jednak všech možných filtrací, které velmi zpřesnily výsledky, a při aplikacích testů byly kritické hodnoty t testu, χ^2 testu a testu poměrem pravděpodobností stanoveny pro $\alpha = 0,005$. Nejvyšší dosažený koeficient úplnosti byl 0,56, což není příliš povzbuzující hodnota. Nabízí se

metoda	poměr detekovaných kolokací
1. slova sousedící	
2. slova sousedící s eliminací stopslov	
3. kolokační okénko	
4. **slova syntakticky závislá**	
5. slova tranzitivně syntakticky závislá	
6. slova v sourozeneckém vztahu	

Tabulka 7.5: Srovnání šesti variant získávání kolokujících bigramů podle průniku objevených kolokací s frekvencí vetší než 5. Výsledky jsou srovnány s metodou 4, která pracuje se syntakticky závislými dvojicemi slov a má největší koeficient přesnosti. Tmavá část ukazuje společně detekované kolokace, světlá část kolokace objevené navíc.

zde tedy otázka, zda lze přesnost výsledků detekce kolokací zlepšit? Možné to je a v následující kapitole ukážeme jak toho docílit. Očekáváme, že čím větší budeme chtít *koeficient úplnosti* R, tím menší bude *koeficient přesnosti* P, a naopak (stejně jako u systémů vyhledávání informací).

7.3.1 Míra kolokativnosti

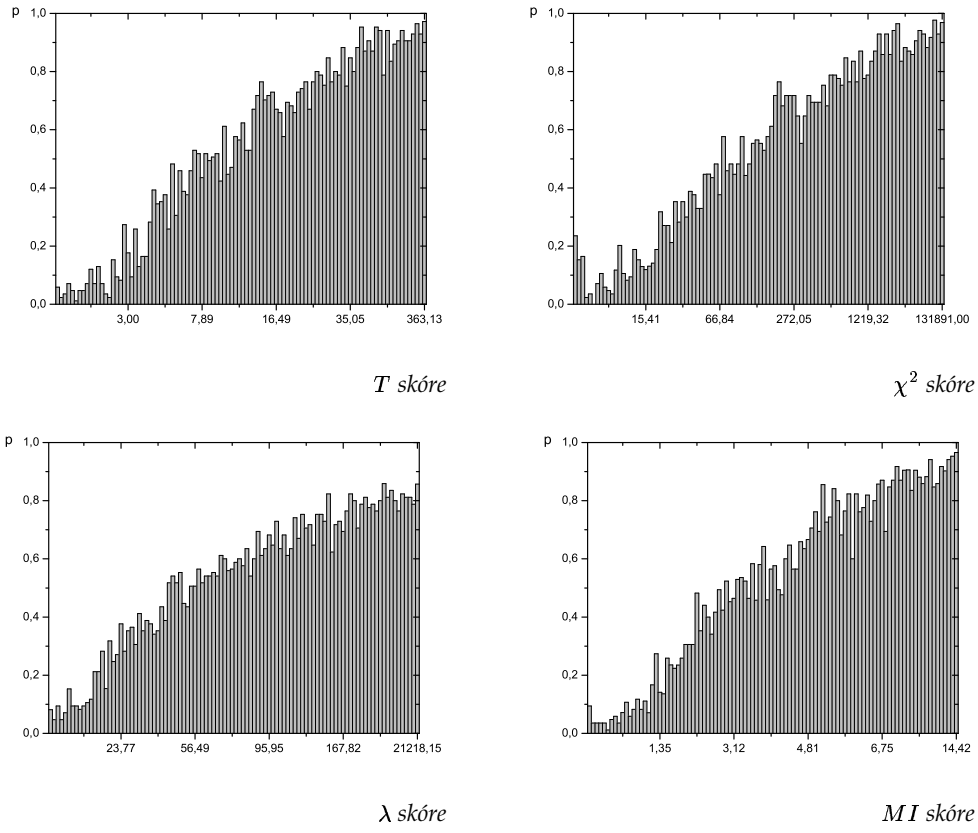
Pro popis *míry kolokativnosti* se nabízejí 4 veličiny:

- *T skóre* - hodnota testové statistiky T použité v *t testu*
- χ^2 *skóre* - hodnota testové statistiky χ^2 použité v χ^2 *testu*
- λ *skóre* - hodnota testové statistiky $-2log\lambda$ použité v *testu*
- MI *skóre* - hodnota vzájemné informace

Naším cílem není vytvoření metody, jejíž výsledkem by byl seznam obsahující pouze a jenom kolokace. Toto totiž ani není možné a vyplývá to již ze subjektivnosti kolokací. Mnohem výhodnější je vytvoření veličiny, která by popisovala *míru kolokativnosti*, tedy „možnost nebo schopnost býti kolokací“. Čím větší by tato *míra kolokativnosti* nějakého bigramu byla, tím větší by byla pravděpodobnost, že se jedná o kolokaci, resp. tím menší by byla pravděpodobnost, že to kolokace není. Pokud budeme mít takovou veličinu k dispozici, jednoduše si podle potřeby zvolíme její kritickou hodnotu. Pokud budeme preferovat přesnost před úplností, bude kritické hodnota větší, v opačném případ ji budeme volit spíše menší.

Na obrázku 7.1 jsou grafy, které dokumentují, na kolik tyto veličiny popisují míru kolokativnosti na seznamu 8 288 bigramů, které se v kolekci $\mathcal{CW}94$ vyskytují jako syntakticky závislé dvojice s frekvencí větší než 5 a které byly ohodnoceny lidskou silou. Grafy vznikly následujícím způsobem: Bigramy jsme setřídili vzestupně podle odpovídajícího skóre. Interval, v němž leží všechny hodnoty této veličiny, jsme rozdělili na 100 podintervalů tak, že každý obsahoval přibližně stejný počet bigramů (82-83). Těchto 100 intervalů odpovídá 100 hodnotám na ose X. Hodnoty na ose Y pak tvoří podíl kolokací v každém z těchto podintervalů (p).

Vhodnost použité veličiny pro měření kolokativnosti pak posuzujeme podle velkého rozptylu hodnot p sousedních intervalů. „Čím menší zuby, tím lépe.“ a podle *vypouklosti* grafu. Ideální tvar



Obrázek 7.1: Chování míry kolokativnosti měřené pomocí hodnot t skóre, χ^2 skóre, $-2log\lambda$ skóre a pomocí hodnot vzájemné informace MI na seznamu kolokací detekovaných v kolekci $\mathcal{CW}94$ jako syntakticky závislé dvojice slov.

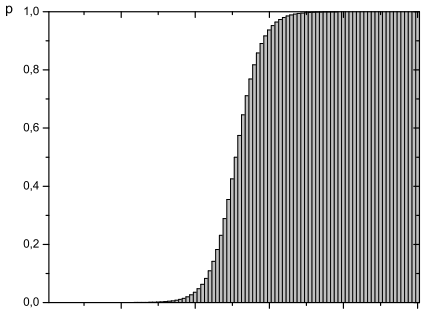
takto vytvořeného grafu je na obrázku 7.2. Pozorujeme, že *T skóre*, χ^2 *skóre* a MI *skóre* popisují kolokativnost přibližně stejně dobře, ale λ *skóre* je výrazně horší. U 82 bigramů s nejvyšším λ *skóre* je podíl kolokací menší než 0,8, u ostatních skóre je to téměř 1,0.

Nyní si položme otázku, zda je možné tyto veličiny nějakým způsobem zkombinovat, tak aby míru kolokativnosti popisovaly lépe. Z několika možných jednoduchých variant se jako nejlepší jeví tyto:

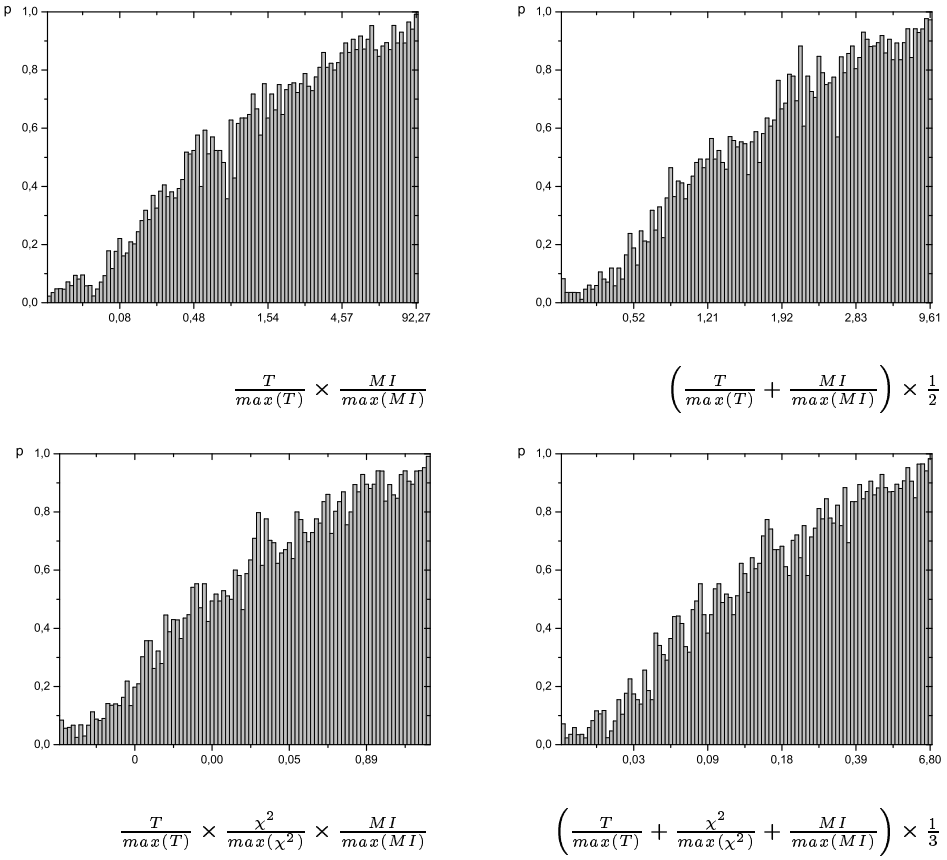
- součin hodnot *T skóre* a *MI skóre* vážených svými maximálními hodnotami $\frac{T}{max(T)} \times \frac{MI}{max(MI)}$
- aritmetický průměr *T skóre* a MI *skóre* vážených svými maximálními hodnotami $\left(\frac{T}{max(T)} + \frac{MI}{max(MI)}\right) \times \frac{1}{2}$

Jejich grafy (a také dvou dalších veličin) jsou na obrázku 7.3. Je zřejmé, že kombinace dvou veličin ke zlepšení vede, ale kombinace tří veličin již očekávané zlepšení nepřináší.

Na závěr podotkněme, že je pravděpodobně možné vytvořit pomocí *T skóre*, MI *skóre* a χ^2 *skóre* složitější veličiny, které mohou míru kolokativnosti popisovat lépe. K jejich odhalení je však



Obrázek 7.2: Ideální chování míry kolokativnosti na seznamu detekovaných kolokacích. Čím strmější a ostřejší „S“ uprostřed tím lepší.



Obrázek 7.3: Chování míry kolokativnosti měřené pomocí váženého součinu T, MI a T, χ^2, MI a pomocí váženého aritmetického průměru T, χ^2, MI na seznamu kolokací detekovaných jako syntakticky závislé dvojice slov v kolekci $\mathcal{CW}94$.

potřeba důkladná analýza dat a také sestavit funkci, která bude měřítkem toho, do jaké míry daná veličina popisuje míru kolokativnosti bigramů.

7.4 Sestavení slovníku kolokací

Všechny experimenty popsané v kapitole 6 jsme provedli na třech různých kolekcích dokumentů, které jsme měli k dispozici (viz Příloha A). Jedná se o kolekce novinových článků z časopisu *Computerworld* ($\mathcal{CW}94$), *Lidových novin* ($\mathcal{LN}92$) a o korpus *Prague Dependency Treebank* ($\mathcal{PDT}$). Pro každou kolekci jsme sestavili šest seznamů kolokujících bigramů získaných šesti různými způsoby. Pro všechny bigramy z těchto $3 \times 6 = 18$ seznamů jsme výpočtem zjistili hodnoty těchto veličin:

- frekvence výskytu
- střední hodnota vzdálenosti komponent (pouze u metody kolokačního okénka a tranzitivní syntaktické závislosti)
- T skóre
- χ^2 skóre
- λ skóre
- MI skóre

Výsledky těchto experimentů jsou na přiloženém CD.

V předchozí kapitole jsme jako nejlepší metodu detekce kolokací uvedli tu, která pracuje se syntakticky závislými dvojicemi slov, vypočítá pro ně hodnoty T skóre a MI skóre a jejich vážený součin použije jako míru kolokativnosti pro setřídění výsledků.

Požadovaný slovník vytvořme tak, že stanovíme kritickou hodnotu pro míru kolokativnosti a vybereme všechny bigramy s frekvencí minimálně 2, které ji překročí. Bude to tedy nějaká úvodní část setříděného seznamu bigramů. Jak lze tuto kritickou hodnotu určit? Řekněme, že požadujeme hodnotu koeficientu přesnosti $P \approx 0,8$. Ze seznamu kolokací s frekvencí větší než 5 ohodnocených lidskou silou zjistíme, že kritická hodnota pro vážený součin T skóre a MI skóre je 0,004. Velikosti výsledných seznamů v porovnání s dalšími hodnotami jsou v tabulce 7.6. Takto vytvořené kompletní slovníky jsou opět na přiloženém CD.

7.5 Poloautomatická detekce kolokací

Zajímavým výsledkem, ke kterému jsme došli během analýzy výsledků experimentů, je princip *poloautomatické detekce kolokací*. V kapitole 2.4 jsme zmiňovali využití kolokací v oblasti komputační lingvistiky a oblasti vyhledávání informací. Rozdíl v nich je ten, že v prvním případě je cílem vytvořit slovník kolokací bez jakýchkoliv chybných a nevhodných hesel. V druhé oblasti tento požadavek nutný není a můžeme se spolehnout na kolokace automaticky detekované s jistým procentem nepřesností. Při vytváření slovníku je však potřeba seznam kolokací zkontrolovat lidskou silou. A právě zde se nabízí použít různých atributů bigramů a výsledků popsaných

	$\mathcal{CW}94$		$\mathcal{PDT}$		$\mathcal{LN}92$	
různá slova	68 907		82 834		223 305	
různá autosémantická slova	49 772	72 %	82 045	99 %	173 445	77 %
různé autosémantické bigramy	251 537		265 180		1 092 966	
různé objevené kolokace	26 851	10 %	32 294	12 %	163 695	15 %

Tabulka 7.6: Souhrn výsledků aplikace detekce kolokací na textové kolekce $\mathcal{CW}94$, $\mathcal{PDT}$, $\mathcal{LN}92$. V prvním řádku jsou počty základních slovních tvarů, na druhých řádcích jsou počety autosémantických základních slovních tvarů. Na dalších pak počet autosémantických bigramů. V posledním řádku jsou pak počty kolokací získaných naší metodou.

metod k předtřídění a uspořádání seznamu kolokujících bigramů, tak že lidem, které budou provádět kontrolu, bude práce co nejvíce usnadněna. Tento princip je již zmíněn v kapitole 6.2, kde jsme popisovali aplikaci pro prohlížení výsledků a mezivýsledků metod detekce kolokací. Jako nejlepší řešení se jeví odstranění všech bigramů, které nevyhovují filtrům popsaným v kapitole 5.6, a setřídění podle míry kolokativnosti pomocí např. *T skóre* nebo součinu hodnot *T skóre* a *MI skóre*, který byl popsán v předchozí kapitole.

Kapitola 8

Zhodnocení praktických výsledků

V této kapitole se budeme stručně zabývat systémy vyhledávání informací v textech a vyzdvihneme přínos kolokací, které jejich využití t této oblasti přináší. Zamyslíme se také nad jejich aplikovatelností a přínosem v praxi. Závěrem diskutujeme míru zastoupení různých typů kolokací.

8.1 Systémy vyhledávání informací

Softwarové systémy s databází textových dokumentů v přirozeném jazyce určené pro ukládání a vyhledávání textů nebo informací v nich nazýváme *dokumentografické informační systémy* (DIS). Dokumentografické informační systémy zpracovávají dokumenty a vytváří speciální datové struktury, které umožňují efektivní vyhledávání. Na textové dokumenty lze pohlížet jako na posloupnosti termů - základních zpracovávaných jednotek. Nejčastěji se jedná o jednotlivé slovní tvary.

Uživatel pokládá systému tzv. *dotazy*, které indikují jeho informační potřebu. Podle typu odpovědi se rozlišují tři typy DIS:

- *Information retrieval* (IR) systémy - Úkolem je vybrat podmnožinu dokumentů, které odpovídají vloženému dotazu.
- *Information extraction* (IE) systémy - Úkolem je nejen vybrat dokumenty, ale také z nich získat požadovanou informaci a formou strukturovaného záznamu ji předat uživateli.
- *Question-answering* (QA) systémy - Cílem je, na položenou otázku v dokumentech najít odpovědi obsahující požadovanou informaci, tuto informaci z textů extrahovat a prezentovat ji uživateli v přirozeném jazyce.

U IR systémů tedy stačí informaci detekovat a označit dokument, ve kterém je obsažena. U IE systému je nutné informaci i extrahovat a podat ji nezávisle na dokumentu (dokumentech), ve kterém (kterých) se nachází. V posledním případě je navíc nutné tuto informaci prezentovat v přirozeném jazyce. Je tedy zřejmé, že náročnost úkolů DIS je nejnižší u IR systémů. IE a QA

systémy ke své práci potřebují daleko detailnější a komplexnější lingvistické znalosti. Pochopení textu, resp. otázky je často složité i pro člověka; stejně tak formulace odpovědí.

Nadále se budeme zabývat právě *Information Retrieval* systémy. Všechny poznatky lze však použít i při studiu vyšších typů *dokumentografických informačních systémů*.

8.1.1 IR systémy

Tradiční IR systémy vyhodnocují dotaz a sestavují odpovědi podle shody slov (termů) v položeném dotazu a v jednotlivých dokumentech. Tento princip však není příliš efektivní, vybrané relevantní dokumenty nemusí slova obsažená v dotazu obsahovat. Důvody, proč může tato situace nastat, mohou být různé.

- *Kritérium predikce* - Jak najít nejvhodnější termy, kterými jsou popsány relevantní dokumenty?
- *Synonyma* - Slova se stejným významem. Jak vědět, které ze slov je v dokumentu použito?
- *Homonyma* - Jak lze u slova s několika různými významy specifikovat význam, který máme na mysli?

Obecné řešení všech těchto problémů je v přechodu od práce se slovy k práci s jejich významy, případně s významy jejich kombinací. K tomuto kroku je nutná lingvistická analýza textu. (Lingvistické znalosti jsou obecně velmi bohaté a rozsáhlé, k jejich automatickém získávání z textů v přirozeném jazyce se v oblasti zpracování přirozeného jazyka (*Natural language processing* - NLP) používají různé tzv. *učící se* techniky.)

IR systémy nepracují s celými dokumenty, nýbrž s jejich reprezentacemi. Je to z důvodu efektivního vyhledávání při zpracování dotazů (příkladem je index nebo soubor vektorů reprezentující dokumenty) Tato reprezentace vždy znamená zmenšení informace, která se o dokumentu uchovává. Podle množství informace uchovávané pro jednotlivé dokumenty lze dle [20] rozdělit IR systémy do 3 úrovní:

1. Systémy, které pracují pouze s informací, zda dokument obsahuje konkrétní term či nikoliv.
2. Systémy, které berou v úvahu i frekvence termů a jejich pozice v dokumentech.
3. Systémy, které analyzují a využívají kontexty jednotlivých výskytů slov (včetně syntaktických vztahů mezi nimi).

První kategorii odpovídá jednoduchý boolský model [21]. Vektorový model [21] lze zařadit do kategorie druhé; váha jednotlivých termů závisí na jejich frekvencích. V obou případech se však jedná o velkou ztrátu informace obsažené v dokumentech. Cílem je tuto informaci zachovat maximální, zároveň také minimalizovat nároky na uložení a rychlost zpracování dotazu a samozřejmě se vyhnout výše zmíněným nedostatkům těchto dvou typů IR systémů. Řešením je extrahovat informace o obsahu dokumentu a o významu jednotlivých slov a jejich spojení. Takové systémy řadíme do třetí kategorie. Právě v nich nalézají kolokace velké uplatnění. Kolokace, jak bylo uvedeno v kapitole 2.3, velmi často vytváří nový význam (nekompozičnost), specifikují význam jednotlivých komponent nebo se jedná časté a ustálené spojení, které může lépe sloužit k určení konkrétního významu slova.

8.1.2 Využití kolokací v IR systémech

Indexování víceslovných termů

Pro tradiční IR systémy je základní zpracovávanou jednotkou (*termem*) slovní tvar, v lepším případě pak lemma. Pokud budeme mít k dispozici seznam kolokací nebo metodu, jak je z dokumentů získat, můžeme tyto víceslovné termy použít stejným způsobem jako termy jednoslovné. V případě, že budou součástí dotazu, vyhledáme dokumenty pouze s výskyty příslušné kolokace.

Uveďme příklad rozšíření jednoduchého boolského modelu o možnost indexovat i kolokace.

Příklad: V kolekci $\mathcal{CW}94$ budeme hledat dokumenty související s pojmem *tabulkový procesor*.

- Slovo *tabulkový* se vyskytuje v 79 dokumentech.
- Slovo *procesor* se vyskytuje v 481 dokumentech.
- Slova *tabulkový* a *procesor* se společně vyskytují v 28 dokumentech.
- Kolokace *tabulkový procesor* se vyskytují v 22 dokumentech.

IR systém s klasický boolským modelem na dotaz *tabulkový procesor* odpoví seznamem 28 dokumentů, které obsahují jak slovo *tabulkový* tak slovo *procesor*. Pouze 22 z nich je pravděpodobně relevantních, protože obsahují celé slovní spojení *tabulkový procesor*. *Koeficient úplnosti* pro tento dotaz je $R = \frac{22}{22} = 1$ a *koeficient přesnosti* $P = \frac{22}{28} = 0,78$.

IR systém, který indexuje navíc i kolokace, vybere přesně 22 dokumentů. které obsahují toto slovní spojení. Předpokládáme, že všechny jsou relevantní, protože tento pojem obsahují. Přesnější údaje o relevantnosti dokumentů nejsme na této úrovni schopni zjistit. *Koeficient úplnosti* i *koeficient přesnosti* pro tento dotaz jsou $R = \frac{22}{22} = 1 = P$. Zlepšení je zřejmé.

Při bližší analýze dokumentů, jsme však zjistili, že zmiňovaných 22 dokumentů opravdu pojednává o *tabulkových procesorech*. Avšak 2 z 6 dokumentů, které kolokaci *tabulkový procesor* neobsahují, se zabývají *tabulkovými kalkulátory*, což je ekvivalentní označení pro *tabulkové procesory*. Skutečně relevantních dokumentů je tedy nejméně 24. *Koeficienty úplnosti* tedy klesne na $R = \frac{22}{24} = 0,92$. Vidíme, že pokud bychom chtěli objevit i tyto dokumenty, musíme opravdu pracovat s významy slov.

Reprezentace nových významů slov

V IR systémech, které pracují s významy slov a nikoli se slovy samotnými, mohou být kolokace použity jako nové významové prvky. Významy lze přitom reprezentovat v IR systémech různě. V případě, že máme k dispozici slovník významů, každému slovu v dokumentu je přiřazen jeho význam z tohoto seznamu (např. slovník *WordNet* [34]). Pokud kolokace v tomto slovníku obsažená není, lze ji do něj vložit jako nový prvek a pracovat s ním stejně.

Významy slov lze také reprezentovat automaticky vytvořenými strukturami. Použitím speciálních algoritmů můžeme sestavit množiny sémanticky podobných slov. Význam každého slova je pak dán množinou sémanticky podobných slov, kterých je prvkem. Kolokace pak můžeme také použít pro vytváření nových významů.

Příklad: Nové významy vzniklé modifikací slova *počítač* v kolekci $\mathcal{CW}94$:

počítač *osobní, PC, sálový, přenosný, hostitelský, střediskový, stolní, Amiga, uzlový, IBM, nadřazený, personální, zápisníkový, multiprocesorový, kompatibilní, centrální, palubní, Apple, paralelní, unixovský, vzdálený, domácí, Cray, značkový, AT, střední, kapesní, číslicový, použitý, multimediální, univerzální, řídicí, hlavní.*

Sémantická disambiguace v dokumentech

Sémantická disambiguace je jedním ze základních problémů NLP. Jejím úkolem je slovu nebo skupině slov v textu přiřadit sémantickou reprezentaci. Kolokace opět tento proces velmi zjednoduší. Kolokace totiž velmi přesně určují význam svých komponent a zároveň nový význam vytváří. Obé bude tedy velmi dobře rozpoznáno, pokud se kolokaci v textu objeví. Získáme tak přesnější významy, které se stanou předmětem indexace.

Příklad: Různé kolokace slova *oko* v korpusu $\mathcal{PDT}$:

{*na*} *vlastní oči, pouhým okem, vystřelené oko, profesionální oko, volské oko, prostým okem, psí oko, Boží oko, Ježíšovo oko, rozzářenýma očima, andělské oči, červené oči.*

Sémantická disambiguace v dotazech

Disambiguaci můžeme použít nejen na dokumenty, ale také na vkládaný dotaz. Systém musí také určit význam tohoto dotazu. Pokud je jeho součástí slovo mnohovýznamové, nemusí být jeho význam ani z kontextu dotazu zřejmý. V tom případě použijeme seznam významů (včetně kolokací) a nabídneme uživateli upřesnění jím míněného významu předložením všech eventualit, které term může mít.

Příklad: Možné významy slova *koruna* v uživatelských dotazech:

koruna *slovenská, česká, dánská, švédská, československá, královská, každá, federální, zlatá, norská, státní, proslulá, směnitelná, tvrdá.*

8.2 Aplikovatelnost

Přesvědčili jsme se, že význam kolokací v *dokumentografických systémech* není nezajímavý. Jak lze ale metody detekce kolokací v této oblasti aplikovat? Mohou nastat tři případy použití:

1. Seznam kolokací je obsažen v rozsáhlém slovníku. Není nutné hledat v dokumentech kolokace nové. Předpokládáme, že všechny jsou ve slovníku a pracujeme pouze s nimi.
2. Kolokace extrahujeme z kolekce dokumentů popsanou metodou. Předpokládáme, že se dokumenty nemění a celý proces vyhledávání kolokací lze provést jednorázově.
3. Kolokace je nutné získat z kolekce zpracovávaných dokumentů, která je velmi proměnlivá a nelze popsanou metodu provádět při každé její změně.

	$\mathcal{CW}94$		$\mathcal{PDT}$		$\mathcal{LN}92$	
všechna slova	1 617 845		1 255 386		8 462 967	
autosémantická slova	912 218	56 %	788 148	62 %	5 187 825	61 %
autosémantické bigramy	481 996		424 091		2 507 436	
detekované kolokace	142 833	30 %	127 932	30 %	1 090 460	43 %
různá slova	68 907		82 834		223 305	
různá autosémantická slova	49 772	72 %	82 045	99 %	173 445	77 %
různé autosémantické bigramy	251 537		265 180		1 092 966	
různé objevené kolokace	26 851	10 %	32 294	12 %	163 695	15 %

Tabulka 8.1: Tabulka výsledků detekce kolokací v textových kolekcích $\mathcal{CW}94$ $\mathcal{PDT}$ $\mathcal{LN}92$ Bigramy jsou myšleny jako slova sousedící v povrchovém slovosledu.

První dva případy jsou bezproblémové, buď se automatická detekce kolokací vůbec neprovádí nebo lze použít naši metodu. V posledním případě však nastává problém, který lze řešit dvěma způsoby:

Periodické zpracování

Detekce kolokací můžeme provádět periodicky po delších časových intervalech. V případě, že v tomto intervalu přibude v kolekci nový dokument, nové kolokace se v něm nehledají, pouze se indexují kolokace již nalezené. Pokud se v něm nalézá nová kolokace vhodná pro indexování, bude objevena až *v dalším kole*. Do té doby zůstane nezpracována.

Iterativní zpracování

Detekce kolokací se provádí průběžně - *iterativně*. Všechny metody používají pro své výpočty údaje o frekvencích výskytů jednotlivých slov, jejich dvojic a také počet všech slov v kolekci. Systém pracuje v každém okamžiku jen s těmi daty, které má k dispozici. Tedy s údaji, které získá z již zpracovaných nebo právě zpracovávaných dokumentů. Systém si všechny potřebné údaje udržuje v paměti a v případě aplikace nějaké metody je použije. Pokud se kolekce dokumentů změní (např. přibudou nové), tak se tyto údaje aktualizují pro slova a bigramy obsažené v nových dokumentech a na dvojice slov z nových kolokací se aplikují výpočty.

V předchozím přístupu se periodicky provádí výpočty pro *všechny* kolokující bigramy. Vybrané kolokace se potom indexují. Iterativní metoda aplikuje výpočty na kolokující bigramy *obsažené v nových dokumentech* a zjistí, zda je potřeba je přidat do seznamu indexovaných termů (kolokací) nebo je z něj naopak vyřadí. Zpočátku nebudou výsledky přesné, ale s narůstající velikostí zpracovaných textů se budou zlepšovat.

Příklad: Průběh iterativní metody

1. Do DIS je na počátku vloženo 1000 dokumentů. Aplikace metod detekce kolokací rozhodne, že 10000 bigramů je vhodné indexovat.
2. Dále jsou do DIS vkládány nové dokumenty, vždy v nějakém časovém intervalu, řekněme 100 denně.

typ kolokace	zastoupení
AN	53,48 %
NN	33,05 %
VN	10,41 %
CN	1,15 %
DA	0,58 %
DV	0,40 %
NA	0,29 %
VN	0,27 %
NC	0,13 %

Tabulka 8.2: Poměr zastoupení jednotlivých slovnědruhových typů kolokací detekovaných v kolekci $\mathcal{CW}94$

3. Každý den se provedou výpočty pro bigramy obsažené v nových dokumentech, tak jako by by byly součástí celé kolekce (první den tedy 1100 dokumentů).

4. Výsledkem může být, že se množinu indexovaných kolokací je nutné rozšířit o nově objevené kolokace (případně z ní vyjmout dvojice o kterých jsme zjistili, že kolokacemi nejsou).

Dále si všimněme nárůstu množství zpracovávaných dat při přechodu od jednoslovných ke dvouslovným termům. V tabulce 8.1 je uveden přehled výsledků metod detekce kolokací v jednotlivých použitých kolekcích.

Zaměřme se na spodní část tabulky. Pro každou kolekci (korpus) jsou zde uvedeny informace o počtu různých slov (základních tvarů slov), které se v nich vyskytují, a také počty různých autosémantických slov a jejich podíl ke slovům všech slovních druhů. Autosémantická slova jsou velmi vhodná pro indexování v IR systémech. Pokud budeme chtít jednoduše rozšířit IR systém o práci se dvojicemi slov, můžeme indexovat všechny bigramy složené ze autosémantických komponent. Počet takových bigramů je uveden v dalším řádku tabulky. Vidíme až několikanásobné navýšení zpracovávaných jednotek($3-5\times$). Pokud však budeme moci použít vyhledávání kolokací ve zpracovávaných dokumentech a rozhodneme se indexovat pouze je, tak se počet těchto dvouslovných termů sníží na hodnoty uvedené v posledním řádku.

8.3 Zastoupení různých druhů kolokací

Další zajímavou skutečností je poměr různých typů kolokací v textech. Na lidskou silou ohodnocených kolokacích (viz kapitola 7.2) pozorujeme následující: Idiomatické výrazy se vyskytují velmi řídce. Slova významově příbuzná v kategorii 1 jsou také velmi vzácná, je to dáno tím, že často je mezi komponentami přece jen syntaktický vztah a jsou zařazeny do kategorie 2. V ní jsou zcela kompoziční kolokace, které jsou však velmi časté, jedna z komponent je obecně velmi často rozvita druhou. Kategorie 4 je častější z důvodu zahrnutí vlastních jmen. Nejčastější je zastoupení nekompozičních kolokací kategorie 3. Tvoří více než dvě třetiny všech kolokací.

Na závěr uveďme, k jakým účelům lze různé typy kolokací použít. Pokud se jedná o tradiční IR systémy, které nepracují s významy slov (úroveň 1 a 2), lze *všechny typy kolokací druhého druhu použít pro indexování*. V IR systémech, které pracují se sémantickými obsahy dokumentů (úroveň 3) je použití následující: Veškeré kompoziční kolokace vytvářejí nový význam a je vhodné je zahrnout do sémantického slovníku a pracovat s nimi jako s významovými termy. Kategorii 2 je možné použít pro rozpoznávání významů jednotlivých komponent. Kategorii 1 pak pro hledání sémanticky podobných slov a automatickou reprezentaci významů slov pomocí jejich kontextů.

Kapitola 9

Závěr

Pokud je nám známo, dosud neexistovala práce, která by se komplexně zabývala problémem automatické detekce sémanticky signifikantních kolokací (dále jen kolokací) v českých textech. V této práci jsme se soustředili na kolokace dvouslovné.

Při pokusech o jejich automatickou detekci jsme vycházeli ze syntézy několika základních statistických metod: měření frekvence slov a bigramů, odhad střední hodnoty a rozptylu vzdálenosti slov a testování hypotéz nezávislosti výskytů slov t testem, χ^2 testem, poměrem pravděpodobností a počítáním vzájemné informace výskytu slov v textu. Ukázalo se být velmi významné, že příprava dat pro statistické zpracování zahrnovala jejich automatickou morfologickou a syntaktickou analýzu.

Hlavní výsledky dosavadní práce lze shrnout takto:

1. **Definice a vlastnosti kolokací.** Novým způsobem jsme definovali sémanticky signifikantní kolokace prvého a druhého druhu. Na základě diskuse o charakteristických vlastnostech kolokací jsme navrhli pětistupňovou klasifikaci kolokací ze sémantickéého hlediska.

2. **Metodologie automatické detekce kolokací.** Metoda automatické detekce kolokací, kterou jsme navrhli, pracuje ve dvou fázích: nejprve z rozsáhlého vzorku textových dat extrahujeme soubor kolokujících bigramů (tj. potenciálních kolokací), a pak tento soubor statisticky analyzujeme. Navrhli jsme šest různých způsobů extrakce kolokujících bigramů, nejen na základě znalosti povrchového slovosledu, ale také s využitím známých syntaktických vztahů ve větách, a zkoumali jsme šest statistických metod pro rozpoznávání kolokací v souboru kolokujících bigramů. Těchto šest metod extrakce bigramů a šest metod výběru kolokací jsme kombinovali a navrhli jsme způsob hodnocení jejich výsledků. Dalšího výrazného zpřesnění výsledků jsme dosáhli filtrací kolokujících bigramů, které kolokace tvořit nemohou.

3. **Experimenty.** Pro výzkumné účely jsme vytvořili aplikaci, která komfortním způsobem umožňuje analyzovat zpracovávaná data. Pro evaluaci testovaných metod a jejich kombinací jsme použili soubor cca 12 000 ručně ohodnocených bigramů. Popsané metody automatické detekce kolokací jsme aplikovali na tři rozsáhlé kolekce českých textů a sestavili jsme tři slovníky sémanticky signifikantních kolokací v rozsahu 26 851 (kolekce $\mathcal{CW}94$), 32 294 (kolekce $\mathcal{PDT}$) a 163 695 (kolekce $\mathcal{LN}92$) dvojslovných hesel.

4. **Aplikace kolokací.** Ukazujeme, že výsledky této práce by měly být prakticky aplikovatelné v oblasti automatického vyhledávání informací v českých textech, a že použití námi navrhovaných metod by mělo významně ovlivnit kvalitu vyhledávání.

Za obzvlášť významné považujeme, že analýzou empirických dat jsme dospěli k důležitému poznatku, že statisticky založená detekce kolokací je úspěšnější, pokud využíváme automatickou analýzu syntaktických závislostí. V další práci se chceme zaměřit zejména na zobecnění metod tak, aby bylo možné detekovat také kolokace více než dvouslovné.

Literatura

[1] Hajič J, Hajičová E., Pajas P., Panevova J., Sgall P., Vidová-Hladká B. (2001): Prague Dependency Treebank 1.0 , *The Linguistic Data Consortium.*

[2] Firth J. R. (1957): A synopsis of linguistic theory 1930-1955, *Oxford: Philological Society.*

[3] Choueka Y. (1988): Looking for needles in a haystack or locating interesting collocational expressions in large textual databases, *In: Proceedings of the RIAO.*

[4] Church Kenneth, William Gale, Patrick Hanks, Donald Hindle (1991): Using statistics in lexical analysis, *In: Uri Zernik (ed.): Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, Lawrence Erlbaum, Hillsdale, NJ.*

[5] Smadja F. (1993): Retrieving collocations from text: Xtract, *In: Computational Linguistics 19.*

[6] Manning C. D., Schütze H. (1999): Foundations of Statistical Natural Language Processing, *MIT Press.*

[7] Sinclair J. (1995): Collins Cobuild English Distionary, *Athelstan Pubns.*

[8] Kocek J., Kopřivová M., Kučera K. (2000): Český národní korpus. Úvod a příručka uživatele, *Ústav českého národního korpusu, FF UK, Praha.*

[9] Strzalkowski T. (Ed.) (1999): Natural Language Information Retrieval, *Kluwer.*

[10] Nešetřil J. (1979): Teorie grafů, *SNTL Praha.*

[11] Justenson John S., Katz Slava M. (1995): Technical terminology: some lingustic properties and an algorithm for identification in text, *In: Natural Language Engineering 1.*

[12] Anděl J. (1998): Statistické metody, *Matfyzpress, Praha.*

[13] Church Kenneth W., Mercer Robert L. (1993): Introduction to the special issue on computational linguistics using large corpora, *In: Computational Linguistics 19, 20.*

[14] Snedecor Georg Wadell, Cochran William (1989): Statistical Methods, *Ames: Iowa State University Press.*

[15] Dunning Ted (1994): Accurate methods for the statistics of surprise and coincidence, *In: Computational Linguistics 19.*

[16] Mood Alexander M., Franklin A. Graybill, Duane C. Boes (1974): Introduction to the theory of statistics, *New York: McGraw Hill.*

[17] Lin D. (1998): Extracting Collocations from Text Corpora, *Computerm '98, Montreal.*

[18] Fano Robert M. (1961): Transmission of Information; A Statistical Theory of Communication, *New York: Wiley.*

[19] Alshawi Hiyan (1994): Training and scaling preference functions for disambiguation, *In: Computational Linguistics 20.*

[20] Holub M. (2000): Context Information Mining I., Motivation and Syntactic Issues, *In: The Prague Buletin of Mathematical Linguistics, UK, Praha.*

[21] Kopecký M. (1998): Dokumetografické informační systémy, *Skripta pro výuku na MFF UK.*

[22] Viegas E. (ed.) (1999): Breath and Depth of Semantic Lexicons, *Kluwer.*

[23] Holub M., Míka P. (2001): Mates — An Experimental Linguistic Database System, *In: Proceedings of the IRCS Workshop on Linguistic Databases, Univesity of Pennsylvania, Philadelphia, USA.*

[24] Pala K., Rychlý P. (1995): Mutual Information in Czech Corpus ESO, *In: Proceedings of the First Workshop on Text, Speech and Dialogue '98. Faculty of Informatics, Masaryk University, Brno.*

[25] Popelínský L., Hanák F. (2001): Objevování znalostí a datamining, *In: Datakon 2001, Proceedings of the Annual Conference.*

[26] Šmilauer V. (1982): Nauka o českém jazyku, *Státní pedagogické nakladatelství, Praha.*

[27] Collins M., Hajič J., Ramshaw L., Tillmann C. (1999): A Statistical Parser for Czech, *ACL '99, Maryland, USA.*

[28] Hajič J., Vidová-Hladká B. (1998): Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset, *Proceedings of the Conference COLING–ACL '98, Montreal, Canada.*

[29] Hajič J. (2000): Morphological Tagging: Data vs. Dictionaries, *6th ANLP Conference, 1st NAACL Meeting. Proceedings, pp. 94–101, Seattle, Washington.*

[30] Hajič J., Krbec P., Květoň P., Oliva K., Petkevič V. (2001): Serial Combination of Rules and Statistics: A Case Study in Czech Tagging, *In: Proceedings of the ACL '2001, Toulouse, France.*

[31] Sgall P. a kolektiv (1986): Úvod do syntaxe a sémantiky, *Academia, Praha.*

[32] Petkevič V. (in press): Underlying Structure of Sentence Based on Dependency, *FF UK, Praha.*

[33] Pecina P. (2002): Sémanticky signifikantní kolokace a jejich význam při vyhledávání informací, *Diplomová práce, MFF UK, Praha.*

[34] Vossen P. (1997): EuroWordNet: a multilingual database for information retrieval, *In: Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, Zurich.*

Příloha A

Zpracovávané kolekce textů

A.1 Popis

Pro výpočty a experimenty jsme měli k dispozici celkem 3 kolekce textů:

$\mathcal{CW}94$

Kolekce $\mathcal{CW}94$ obsahuje články z české verze časopisu Computerworld z roku 1994, anotované automatickým lingvistickým parserem, obsahovaly i informace o syntaktické stavbě vět, nikoli však hodnoty analytické funkce. Obsahuje 1430 dokumentů. Zaměření časopisu (informační technologie) výrazně ovlivňuje jeho obsah, resp skladbu použitých slov a především kolokace a jejich tématický okruh. Průměrná úspěšnost rozpoznání slovních tvarů je lehce zvýšená díky množství anglických slov a zkratek. Tento korpus obsahoval asi 6 % nevhodných dokumentů. Většinou se jednalo o články ve slovenštině nebo měly dokumenty jiné kódování češtiny než ISO Latin 2.

$\mathcal{PDT}$

Korpus $\mathcal{PDT}$ je součástí tzv. *Prague Dependency Treebank*. Jedná se o ručně anotované texty nejrůznějšího původu. Celkem ve 5860 dokumentech. Obsahuje jak syntaktické závislosti, tak hodnoty analytické funkce. Úspěšnost správného rozpoznání slovních tvarů je 100 %. S touto přesností jsou rozpoznány také syntaktické závislosti a hodnoty analytické funkce.

$\mathcal{LN}92$

Kolekce $\mathcal{LN}92$ obsahuje články Lidových novin z roku 1992. Jedná se o 34493 dokumentů z nejrůznějších oblastí. Vždy však mají charakter novinových zpráv, komentářů, fejetonů apod. Průměrná úspěšnost rozpoznávání slov není nijak zvýšená. Poměr nevhodných dokumentů je méně než 1 %, jsou to především dlouhé seznamy sportovních výsledků apod.

A.2 Statistika

Kolekce

kolekce	$\mathcal{CW}94$	$\mathcal{PDT}$	$\mathcal{LN}92$
dokumentů	1 413	5 860	34 493
nevyhovujících dokumentů	86	0	260
vyhovujících dokumentů	1 327	5 860	31 833
odstavců	14 342	38 233	128 585
vět	92 672	81 601	487 912
všech slovních tvarů	1 617 846	1 255 385	8 462 967
různých slovních tvarů	137 048	172 472	483 573
různých základních slovních tvarů	68 907	82 834	223 305
všech nerozpoznaných slovních tvarů	67 716	43	90 810
různých nerozpoznaných slovních tvarů	17 851	20	46 227
všech nerozpoznaných slovních tvarů (%)	4,19	0,00	1,07
různých nerozpoznaných slovních tvarů (%)	13,03	0,00	9,56

Dokumenty průměrně

kolekce	$\mathcal{CW}94$	$\mathcal{PDT}$	$\mathcal{LN}92$
odstavců	10,80	6,52	4,04
vět	69,83	13,92	15,32
všech slovních tvarů	985,77	182,34	224,39
různých slovních tvarů	594,54	143,97	170,13
různých základních slovních tvarů	440,71	122,7	143,88
všech nerozpoznaných slovních tvarů	5,35	0,01	1,52
různých nerozpoznaných slovních tvarů	5,00	0,00	1,63
všech nerozpoznaných slovních tvarů (%)	1,20	0,00	3,11
různých nerozpoznaných slovních tvarů (%)	1,53	0,00	3,42

Slovní druhy

kolekce		$\mathcal{CW}94$		$\mathcal{PDT}$		$\mathcal{LN}92$	
značka	slovní druh	$\sum$	%	$\sum$	%	$\sum$	%
N	podstatná jména	432 481	26,73	379 223	23,44	2 486 802	29,38
A	přídavná jména	167 631	10,36	147 622	9,12	920 839	10,88
P	zájmena	91 113	5,63	81 532	5,04	536 667	6,34
C	číslovky	48 954	3,03	40 343	2,49	361 008	4,26
V	slovesa	184 824	11,42	154 893	9,57	993 051	11,73
D	příslovce	78 328	4,84	66 067	4,08	426 125	5,04
R	předložky	134 866	8,34	121 279	7,50	805 214	9,51
J	spojky	85 137	5,26	70 699	4,36	457 165	5,40
T	částice	5 224	0,32	6 689	0,41	32 923	0,38
I	citoslovce	224	0,01	109	0,01	779	0,01
X	nerozpoznaná	79 334	4,90	58	0,01	122 401	1,44
Z	interpunkce	309 728	19,14	186 871	11,55	1 319 993	15,59

Příloha B

Vybrané SGML značky v anotovaných textech

Při procesu značkování jsou texty pomocí tzv. značek opatřovány doprovodnými informacemi (anotovány). Text se na základě tokenizace člení do hierarchicky uspořádaných úseků. Každý úsek je dle standardu SGML uvozen tzv. *otevírací značkou* ve tvaru `<značka>`, poté následuje příslušný úsek textu, který bývá ukončen tzv. *uzavírací značkou* ve tvaru `</značka>` nebo otevírací značkou nového úseku. Každá otevírací značka může také obsahovat přídavné informace ve tvaru `<značka atribut="hodnota">`.

Podle druhu doprovodných informací, které obsahují, rozdělujeme značky na správní, strukturní a lingvistické.

B.1 Správní značky

Správní značky tvoří tzv. hlavičku, která obsahuje administrativní údaje o textu. Jsou to především informace o původu, autorství, typu, zdroji textu a způsobu značkování. Jsou to nejčastěji tyto:

`<csts>` – jazyk dokumentu
`<h>` – hlavička
`<source>` – původ, zdroj dokumentu
`<markup>` – informace o způsobu značkování
`<doc>` – identifikátor dokumentu
`<a>` – informace o modu, typu, žánru, zda-li je text veršován, druh media, na kterém byl publikován, pohlaví autora, jazyk, rok vydání, 1. rok vydání a pod.

Příklad: Hlavička anotovaného textu

```
<csts lang=cs>
```

```
<h>
<source>Lidové noviny</source>
<markup>
<mauth>Jan Hajic
<mdate>Fri Jan 21 01:38:34 2000
<mdesc>Morphology; parameters: RootOut=0, EndOut=0, AllTags=0,
       LemmaTagUpdate=1, ForceLemmaUpdate=1
<mdesc>Morphology; parameters, set 2: MDCopy=1
<mdesc>MA: syn/sem/sty: :_W_T_B/;_G_Y_S_E_R_K_H_U_L/, _x_s_a_n_h_e_l_v_t/
<mdesc>MA: output: desc: ^Yes, la:/lc:_s_a_n_h_e_l_v, va:/vc:-3-4-5-6-7-8-9;
</markup>
<doc file=”s/inf/j/1994/cmpr9406” id=”027”>
</h>
<a>
<mod>s
<txtype>inf
<genre>mix
<med>j
<temp>1994
<authname>y
<opus>cmpr9406
<id>027
</a>
```

B.2 Strukturní značky

Strukturní značky člení text do hieararchicky uspořádaných úseků, odstavců, vět a tokenů (slov nebo interpunkčních znamének).

- `<p>` – odstavec
- `<s>` – věta
- `<f>` – slovo, ne však interpunkce
- `<d>` – interpunkce
- `<D>` – informace o tom, že na daném místě v textu není mezera

Příklad: Označkovaná věta

```
<p n=24>
<s id=”docID:001-p1s365”>
<f cap>Děkujeme<MDl>děkovat_:T<MDt>VB-P---1P-AA---<r>3<g>2<A>Pred
<D>
```

```
<d>. <MDl>. <MDt>Z:-------------<r>8<g>0<A>AuxK
```

B.3 Lingvistické značky

Lingvistické značky popisují morfologickou interpretaci a syntaktický vztah každého tokenu. Každý řádek anotovaného textu, který začíná značkou `<f>` (slovo) nebo `<d>` (interpunkce) osahuje v uvedených značkách tyto informace:

`<f>` resp. `<d>` – použitý slovní tvar
`<MDl>` – lemma s eventuálním upřesněním významu
`<MDt>` – morfologická značka
`<r>` – index tokenu ve větě, pořadové číslo
`<g>` – index tokenu, na kterém je token syntakticky závislý
`<A>` – druh syntaktické závislosti

Značka `<f>` může být také doplněna následujícími atributy:

`abbr` – zkratka
`cap` – první písmeno velké
`mixed` – složeno z písmen i číslic
`num` – číslo
`upper` – všechna písmena velká

Příklad: Označkované tokeny

```
<f abbr>dr<MDl>Dr_:B<MDt>Xx-------------<r>6<g>8
<f cap>Stejně<MDl>stejně_^(*1ý)<MDt>Dg-------1A----<r>1<g>6
<f mixed>24bitovým<MDl>24bitovým<MDt>X@-------------<r>13<g>10
<f num>150<MDl>150<MDt>C=-------------<r>7<g>5
<f upper>IBM<MDl>IBM_:B_;K_^(International_Business_Machines)
<MDt>NNFXX-----A----<r>13<g>11
```

Příloha C

Popis morfologických značek

Součástí výsledku morfologické analýzy jsou tzv. morfologické značky. Spolu s lemmatem jednoznačně identifikují slovní tvar hesla. V případě syntézy je tedy možné z lemmatu a morfologické značky odpovídající slovní tvar vytvořit.

C.1 Struktura značky

Morfologická značka je řetězcem 15 znaků. Je sestavena tak, aby každá její pozice odpovídala jedné morfologické kategorii. Každé hodnotě v dané kategorii odpovídá jeden znak. Ve většině případů jsou to písmena velké abecedy a číslice, výjimečně i nealfanumerické znaky. Hodnota, která nedává smysl (např. pád u sloves), je reprezentována znakem '-' (pomlčka). Morfologické značky většinou odpovídají tradičnímu lingvistickému pojetí, ale není tomu tak důvodů vždy. V případech nejednoznačnosti nebo také neurčitosti, se používá znak 'X'.

C.2 Popis jednotlivých pozic značky

Pozice 1 - Slovní druh (POS)

Označuje hlavní slovní druh. Většinou odpovídá tradiční české gramatice, vyskytují se však výjimky. Je to zejména v případech, kdy zařazení není jednoznačné nebo se gramatiky a slovníky v určení slovního druhu neshodují.

- A – adjektivum (přídavné jméno)
- C – numerál (číslovka, nebo číselný výraz s číslicemi)
- D – adverbium (příslovce)
- I – interjekce (citoslovce)
- J – konjunkce (spojka)
- N – substantivum (podstatné jméno)
- P – pronomen (zájmeno)
- R – prepozice (předložka)

T – partikule (částice)
V – verbum (sloveso)
X – neznámý, neurčený, neurčitelný slovní druh
Z – interpunkce, hranice věty

Pozice 2 - Detailní určení slovního druhu (SUBPOS)

Detailní slovní druh určující ”podkategorie” hlavních slovních druhů. Vždy je z něj možné hlavní slovní druh jednoznačně odvodit.

! – zkratka jako adverbium
– hranice věty (jen u ”virtuálního” slova ”###”)
– slovo ”krát” (slovní druh: spojka)
, – spojka podřadící (vč. ”aby” a ”kdyby” ve všech tvarech)
. – zkratka jako adjektivum
0 – předložka s připojeným ”-ň” (něj), proň, naň, atd. (slovní druh: zájmeno)
1 – vztažné přivlastňovací zájmeno ”jehož”, ”jejíž”, . . .
2 – pomlčka (vždy je samostatně)
3 – zkratka jako číslovka
4 – vztažné zájmeno s adjektivním skloňováním (obou typů: jaký, který, čí, . . .)
5 – zájmeno ”on” ve tvarech po předložce (tj. ”n-”: něj, něho, . . .)
6 – reflexívní zájmeno ”se” v dlouhých tvarech (sebe, sobě, sebou)
7 – reflexívní zájmeno ”se”, ”si” pouze v těchto tvarech, a dále ”ses”, ”sis”
8 – přivlastňovací zájmeno ”svůj”
9 – vztažné zájmeno ”jenž”, ”již”, . . . po předložce (”n-”:”něhož”, ”níž”, . . .)
: – interpunkce všeobecně (ne však ”virtuální” slovo ### jako hranice věty)
; – zkratka jako substantivum
= – číslo psané číslicemi (slovní druh: ’C’)
? – číslovka ”kolik”
@ – slovní tvar, který nebyl morfologickou analýzou rozpoznán (sl. druh: ’X’)
A – adjektivum obyčejné
B – sloveso, tvar přítomného nebo budoucího času
C – adjektivum, jmenný tvar
D – zájmeno ukazovací (ten, onen, . . .)
E – vztažné zájmeno ”což”
F – součást předložky, která nikdy nestojí samostatně (nehledě, vzhledem, . . .)
G – přídavné jméno odvozené od slovesného tvaru přítomného přechodníku
H – krátké tvary osobních zájmen (mě, mi, ti, mu, . . .)
I – citoslovce (slovní druh: ’I’)
J – vztažné zájmeno ”jenž” (”již”, . . .), bez předložky
K – zájmeno tázací nebo vztažné ”kdo”, vč. tvarů s ”-ž” a ”-s”
L – zájmeno neurčité ”všechen”, ”sám”
M – přídavné jméno odvozené od slovesného tvaru minulého přechodníku
N – substantivum, obyčejné
O – samostatně stojící zájmena ”svůj”, ”nesvůj”, ”tentam”
P – osobní zájmena (vč. tvaru ”tys”)
Q – zájmeno tázací/vztažné ”co”, ”copak”, ”cožpak”
R – předložka, obyčejná

S – zájmeno přivlastňovací ”můj”, ”tvůj”, ”jeho” (vč. plurálu)
T – částice (slovní druh ’T’)
U – adjektivum přivlastňovací (na ”-ův” i ”-in”)
V – předložka vokalizovaná (”ve”, ”pode”, ”ku”, . . .)
W – zájmena záporná (”nic”, ”nikdo”, ”nijaký”, ”žádný”, . . .)
X – slovní tvar, který byl rozpoznán, ale značka (ve slovníku) chybí
Y – zájmeno ”co” spojené s předložkou (”oč”, ”nač”, ”zač”)
Z – zájmeno neurčité (”nějaký”, ”některý”, ”číkoli”, ”cosi”, . . .)
^ – spojka souřadicí
a – číslovka neurčitá (”mnoho”, ”málo”, ”tolik”, ”několik”, ”kdovíkolik”, . . .)
b – příslovce (bez určení stupně a negace; ”pozadu”, ”naplocho”, . . .)
c – kondicionál slovesa být (”by”, ”bych”, ”bys”, ”bychom”, ”byste”)
d – číslovka druhová, adjektivní skloňování (”jedny”, ”dvojí”, ”desaterý”, . . .)
e – slovesný tvar přechodníku přítomného (”-e”, ”-íc”, ”-íce”)
f – slovesný tvar: infinitiv
g – příslovce (s určením stupně a negace; ”velký”, ”zajímavý”, . . .)
h – číslovky druhové ”jedny” a ”nejedny”
i – slovesný tvar rozkazovacího způsobu
j – číslovka druhová ≥ 4, substantivní postavení (”čtvero”, ”desatero”, . . .)
k – číslovka druhová ≥ 4, adjektivní postavení, krátký tvar (”čtvery”, . . .)
l – číslovky základní 1-4, ”půl”, . . . ; sto a tisíc v nesubstantivním skloňování
m – slovesný tvar přechodníku minulého, příp. (zastarale) přech. přít. dokonavý
n – číslovky základní ≥ 5
o – číslovky násobné neurčité (”-krát”: ”mnohokrát”, ”tolikrát”, . . .)
p – slovesné tvary minulého aktivního příčestí (vč. přidaného ”-s”)
q – archaické slovesné tvary minulého aktivního příčestí (zakončení ”-t”)
r – číslovky řadové
s – slovesné tvary minulého pasívního příčestí (vč. přidaného ”-s”)
t – archaické slovesné tvary přitomného a budoucího času (zakončení ”-t”)
u – číslovka tázací násobná ”kolikrát”
v – číslovky násobné (”-krát”: ”pětkrát”, . . .)
w – číslovky neurčité s adjektivním skloňováním (”nejeden”, ”tolikátý”, . . .)
x – zkratka, slovní druh neurčen/neznámý
y – zlomky zakončené na ”-ina” (slovní druh: ’C’)
z – číslovka tázací řadová ”kolikátý”
} – číslovka psaná římskými číslicemi
~ – zkratka jako sloveso

Pozice 3 - Jmenný rod (GENDER)

– – neurčuje se
F – femininum (ženský rod)
H – femininum nebo neutrum (tedy nikoli maskulinum)
I – maskulinum inanimatum (rod mužský neživotný)
M – maskulinum animatum (rod mužský životný)
N – neutrum (střední rod)
Q – femininum singuláru nebo neutrum plurálu (pouze u příčestí a jmenných adj.)
T – masculinum inanimatum nebo femininum (jen plurál u příčestí a jm. adj.)

X – libovolný rod (F/M/I/N)
Y – masculinum (animatum nebo inanimatum)
Z – 'nikoli femininum' (tj. M/I/N; především u příslovcí)

Pozice 4 - Číslo (NUMBER)

– – neurčuje se
D – duál (pouze 7. pád feminin)
P – plurál (množné číslo)
S – singulár (jednotné číslo)
W – pouze v kombinaci s jmenným rodem 'Q' (singulár pro fem., pl. pro neutra)
X – libovolné číslo (P/S/D)

Pozice 5 - Pád (CASE)

– – neurčuje se
1 – nominativ (1. pád)
2 – genitiv (2. pád)
3 – dativ (3. pád)
4 – akuzativ (4. pád)
5 – vokativ (5. pád)
6 – lokativ (6. pád)
7 – instrumentál (7. pád)
X – libovolný pád (1/2/3/4/5/6/7)

Pozice 6 - Přivlastňovací rod (POSSGENDER)

Rody mužský neživotný a střední se nikdy nevyskytují samostatně. 'M' se může vyskytnout jen u přivlastňovacích adjektiv (me u příslovcí).

– – neurčuje se
F – femininum (ženský rod)
M – maskulinum animatum (rod mužský životný)
X – libovolný rod (F/M/I/N)
Z – 'nikoli femininum' (tj. M/I/N; u přivlastňovacích příslovcí)

Pozice 7 - Přivlastňovací číslo (POSSNUMBER)

– – neurčuje se
P – plurál (množné číslo)
S – singulár (jednotné číslo)

Pozice 8 - Osoba (PERSON)

– – neurčuje se

1 – 1. osoba
2 – 2. osoba
3 – 3. osoba
X – libovolná osoba (1/2/3)

Pozice 9 - Čas (TENSE)

– – neurčuje se
F – futurum (budoucí čas)
H – minulost nebo přítomnost (P/R)
P – prézens (přítomný čas)
R – minulý čas
X – libovolný čas (F/R/P)

Pozice 10 - Stupeň (GRADE)

– – neurčuje se
1 – 1. stupeň
2 – 2. stupeň
3 – 3. stupeň

Pozice 11 - Negace (NEGATION)

– – neurčuje se
A – afirmativ (bez negativní předpony ″ne-″)
N – negace (tvar s negativní předponou ″ne-″)

Pozice 12 - Aktívum/pasívum (VOICE)

– – neurčuje se
A – aktívum
P – pasívum

Pozice 13 - Nepoužito (RESERVE1)

– – neurčuje se

Pozice 14 - Nepoužito (RESERVE2)

– – neurčuje se

Pozice 15 - Varianta, stylový příznak apod. (VAR)

– – neurčuje se (”základní” tvar pro kategorie v pozicích 1-14)
1 – varianta, víceméně rovnocenná (”méně častá”)
2 – řídká, archaická nebo knižní varianta
3 – velmi archaický tvar, též hovorový
4 – velmi archaický nebo knižní tvar, pouze spisovný (ve své době)
5 – hovorový tvar, ale v zásadě tolerováno ve veřejných projevech
6 – hovorový tvar (koncovka standardní obecné češtiny)
7 – hovorový tvar (koncovka standardní obecné češtiny), varianta k ’6’
8 – zkratky
9 – speciální použití (tvary zájmen po předložkách apod.)

Příloha D

Přehled hodnot analytické funkce

V následující tabulce jsou uvedeny všechny přípustné hodnoty atributu afun. Všechny se mohou vyskytovat také s "příponou" *_Co* (součást koordinace), resp. *_Ap* (součást aposice), resp. *_Pa* (řídící uzel vsuvky - parentese).

Pred	–	Predikát, resp. uzel, který nezávisí na jiném uzlu; věší se na #.
Sb	–	Subjekt (podmět).
Obj	–	Ojekt (předmět).
Adv	–	Adverbiale (příslovečné určení, bez dalšího rozlišení).
Atv	–	Doplněk (jen tzv. určující), technicky zavěšen na neslovesném členu.
AtvV	–	Doplněk (jen tzv. určující), visící na slovese (chybí druhý řídící člen).
Atr	–	Atribut (přívlastek).
Pnom	–	Predikát nominální, resp. jmenná část přísudku se sponou být.
AuxV	–	Pomocné sloveso být (Auxiliary Verb).
Coord	–	Koordinační uzel (souřadné spojení).
Apos	–	Aposice (hlavní uzel).
AuxT	–	Zvratné se, neoddělitelné se - reflexivní tantum.
AuxR	–	Zvratné se, které není Obj ani AuxT (tvoří pasivum reflexivní).
AuxP	–	Předložka primární, části předložky sekundární.
AuxC	–	Spojka (podřadicí).
AuxO	–	Nadbytečný (odkazovací, emotivní) element.
AuxZ	–	Zdůrazňovací slovo.
AuxX	–	Čárka (ne však nositel koordinace).
AuxG	–	Jiné grafické symboly, které neukončují větu.
AuxY	–	Příslovce a částice, které nelze zařadit jinam.
AuxS	–	Kořen stromu (#).
AuxK	–	Koncová interpunkce věty.
ExD	–	Náhradní funkce pro technické hrany vedoucí místo od elidovaného členu k "pseudořídícímu" slovu nebo pro hlavní člen věty bez predikátu (Ex-Dependent).
AtrAtr	–	Řídícím slovem atributu může být díky strukturní víceznačnosti kterékoli z bezprostředně předcházejících (syntaktických) substantiv.
AtrAdv	–	Strukturnívíceznačnost mezi závislostí adverbální (příslovečnou) a adnominální (zavěšení na jméno) bez sémantických důsledků.
AdvAtr	–	Dtto, s opačnou preferencí.

AtrObj – Strukturní víceznačnost mezi závislostí objektovou a adnominální (zavěšení na substantivum) bez sémantických důsledků.

ObjAtr – Dtto, s opačnou preferencí.