

UNIVERZITA PALACKÉHO

FILOZOFICKÁ FAKULTA

Katedra bohemistiky



Miroslav Kubát

**Rozsah, rozptýlení a koncentrace lexika
v psané a mluvené češtině**

Bakalářská práce

Vedoucí práce: PhDr. Ondřej Bláha, Ph.D.

Olomouc 2010

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně na základě uvedené literatury.

V Olomouci 21. dubna 2010

.....

Rád bych poděkoval PhDr. Ondřeji Bláhovi, Ph.D. za vstřícné vedení mé bakalářské práce.

OBSAH

1	ÚVOD	6
2	DEFINICE PSANOSTI A MLUVENOSTI	8
2.1	K pojmu <i>mluvená čeština</i>	8
2.2	Psanost vs. mluvenost.....	10
2.3	Podstata difference psaných a mluvených textů.....	11
3	METODOLOGIE LEXIKÁLNÍ ANALÝZY	13
3.1	Výběr a rozsah materiálu (korpusu).....	13
3.1.1	Kvantitativní hledisko při vytváření korpusu	14
3.1.2	Kvalitativní hledisko při vytváření korpusu.....	14
3.1.2.1	Psaný korpus	15
3.1.2.2	Mluvený korpus	16
3.2	Lexikální jednotka	16
3.2.1	Negace.....	17
3.2.2	Stupňování adjektiv a adverbíí	17
3.2.3	Zájmena	18
3.2.4	Sloveso <i>být</i>	18
3.2.5	Předložky	18
3.3	Úskalí statistiky	19
3.4	Některé práce zabývající se frekvenční analýzou češtiny	20
3.4.1	Psaná čeština.....	20
3.4.2	Mluvená čeština	21
4	ROZSAH LEXIKA	22
4.1	Definice.....	22
4.2	Výpočet.....	23
4.3	Analýza	23
4.3.1	Připravenost-nepřipravenost	24
4.3.2	Charakter podavatele	25
4.3.3	Téma textů	26
4.3.4	Oficiálnost-soukromost.....	26
4.3.5	Explicitnost a určitost.....	27

4.3.6	Abstrakce	27
4.3.7	Organizovanost	28
5	ROZPTÝLENÍ LEXIKA	29
5.1	Definice	29
5.2	Výpočet	29
5.3	Analýza	30
6	KONCENTRACE LEXIKA.....	31
6.1	Definice	31
6.2	Výpočet	31
6.3	Analýza	32
7	ROZBORY	35
7.1	Zájmeno <i>ten</i>	35
7.2	Výplňková slova	37
7.3	K (ne)dominanci spojky <i>a</i>	40
7.4	Komparace hodnot s FSVS	43
7.5	Fatické prostředky v mluvených textech	45
7.6	Komparace našich slovníků s FSC, FSSDTČJ a FSMČ.....	46
8	ZIPFOVY ZÁKONY	48
8.1	První Zipfův zákon	48
8.2	Druhý Zipfův zákon	52
8.3	Třetí Zipfův zákon	54
9	ZÁVĚR	55
10	ANOTACE.....	58
11	ZDROJE	60
12	PŘÍLOHA	62

1 ÚVOD

Cílem této práce je stanovit a popsat rozdíly psané a mluvené češtiny na základě lexikální analýzy psaného a mluveného korpusu. Zjišťovány budou některé charakteristiky kvantitativní lingvistiky, především půjde o rozsah, rozptýlení a koncentraci lexika. Vedle těchto elementárních charakteristik budou také sledovány další parametry, na nichž budou srovnávána specifika textů z hlediska psanosti a mluvenosti.

Nejprve budou sestaveny dva stejně rozsáhlé korpusy – psaný a mluvený. Následně budou na jejich základě vytvořeny frekvenční slovníky, které jsou pro lexikální analýzu a většinu metod kvantitativní lingvistiky výchozím materiálem. Následně budou hodnoty získané při analýze a výpočtech srovnávány, přičemž bude sledován především rozdíl výše zmíněných charakteristik mezi psanými a mluvenými texty. Rozdíly zjištěných hodnot budou podrobeny rozborům, které budou mít za cíl vysvětlit jak samotnou existenci zjištěných diferencí, tak i jejich markantnost. Výstupy rozborů zjišťovaných parametrů budou konfrontovány s interpretací těchto jevů v odborných publikacích.

Je třeba hned v úvodu zdůraznit, že tato práce v žádném případě nespadá primárně do kvantitativní lingvistiky jako takové, samotné prvotní analýzy zkoumaných textů a jejich zpracování budou pouze podkladem pro následné rozborů a zasazování jejich výsledků do širšího kontextu problematiky psané a mluvené češtiny. Prvotní lexikální analýza tak bude zpracována primárně s ohledem k uvedeným cílům celé práce a nepůjde tak o komplexní lexikální analýzu, nebo dokonce kvantitativní analýzu všech jazykových plánů.

Získané hodnoty lexikální analýzy budou srovnávány s nejdůležitějšími pracemi kvantitativní lingvistiky, zejména půjde o nejnovější frekvenční slovníky psané a mluvené češtiny, ale bude přihlédnuto i ke starším publikacím. Tyto komparace mají dvojitý význam, jednak bude ověřena správnost či nesprávnost námi zvolených metod, zejména při sestavování frekvenčních slovníků, jednak přímou konfrontací obdobných prací bude možné sledovat určité jevy v psané i mluvené češtině.

Část práce bude věnována také tzv. Zipfovým zákonům, jejichž platnost bude ověřována na sestavených frekvenčních slovnících. Sledována bude nejen jejich

platnost pro češtinu jako homogenní systém přirozeného jazyka, ale také bude věnována pozornost korelaci míry platnosti Zipfových zákonů na psanosti a mluvenosti.

Součástí práce je příloha, která obsahuje vytvořené frekvenční slovníky psané a mluvené češtiny, přičemž jsou uvedeny pouze lexikální jednotky s frekvencí 5 a vyšší, a to jednak z kvantitativních důvodů a jednak z důvodu relevance (ve smyslu vypovídací hodnoty charakteristik rozdílů psané a mluvené češtiny).

2 DEFINICE PSANOSTI A MLUVENOSTI

Tato práce již ve svém názvu implikuje, že se zde budeme neustále zabývat konfrontací psané a mluvené češtiny, máme proto za vhodné se hned v úvodní části textu pokusit vymezit samotné pojmy *psanost* a *mluvenost*. Zejména mluvená forma češtiny je aktuálně v centru pozornosti lingvistů, o čemž svědčí řada nových publikací a pořádaných konferencí. Zvýšený zájem o zkoumání právě mluveného jazyka má patrně dvě elementární příčiny. První spočívá v rychlém vývoji technologií, které nám umožňují mluvený jazyk zkoumat, dříve byli lingvisté značně omezováni právě technickými limity, druhá příčina pak tkví ve společenských změnách, kdy schopnost účinné mluvené komunikace znovu nabývá na významu.

2.1 K pojmu *mluvená čeština*

Považujeme za vhodné především výrazu *mluvená čeština* zde věnovat pozornost, neboť vymezení jeho významu může být značně problematické.

Stejný úkol si dal i František Daneš při svém úvodním slovu při zahajování mezinárodní vědecké konference *K diferenciaci současného mluveného jazyka*, která se uskutečnila roku 1994 v Ostravě. Daneš se k tomuto problému vyjádřil takto: „Pokud jde o *mluvený jazyk*, soudím, že organizátoři našeho zasedání měli patrně na mysli to, co bych nejráději nazýval *jazyk běžně mluvený*, tedy každodenní neformální ústní komunikaci. Obecně vzato, pod termín *mluvený jazyk*, pokud bychom jej chápali jako všechny texty (diskurzy) realizované zvukově, by patřily značně různorodé typy textů, v nejrůznějších, velmi odlišných komunikativních situacích. – Pro jistotu poznamenám, že termín *jazyk běžně mluvený* neužívám pro označení nějakého *útvary* (či *variety*) nespisovného, *substandardního*, nýbrž jej chápu situačně, jako množinu promluv v komunikačních situacích, v nichž se nepředpokládá (závazně) užívání spisovného jazyka (tedy o situacích *neformálních*, v kterých podle Hallidaye, jazyková forma nevystupuje do popředí zájmu či pozornosti), a zároveň ovšem repertoár všech různých i různorodých

výrazových prostředků (nespisovných, ale zčásti i spisovných, resp. společných), jichž se v tomto druhu promluv užívá.“¹

Podobně vnímá mluvený jazyk i Jan Chloupek, který mluvené texty (v užším pojetí) definuje jako součásti komunikátů:

- „které nebyly předtím zaznamenány, zapsány;
- které už dříve nebyly určeny k přednášení, referování a které nebyly předem promyšleny;
- které nesledovaly cílevědomý estetický záměr mluvčího a nebyly poetizovány, např. veršem, rýmem, čistě individuální intonací, nadto se nevyskytovaly v pevných celcích;
- které lze považovat za přirozené, tedy tematicky spojené s člověkem, se společností, s přírodou, nikoli s jazykem nebo s úzkou profesní specializací.“²

Chloupkovo vymezení se tak v podstatě ztotožňuje s Danešovým pojetím *jazyka běžně mluveného*.

Jiří Kraus také zmiňuje potřebu přesného vymezení výrazu *mluvená čeština*, kdy chápe tento výraz ve dvou základních rovinách, rozlišuje mluvenost vymezenou pouze na základě fonetické realizace a mluvenost funkční. Upozorňuje také na významný současný trend, kdy dochází k interferenci mluvenosti a psanosti, a to především kvůli nárůstu tzv. elektronické komunikace. Kraus tak považuje projevy čistě psané nebo mluvené jen za teoretické konstrukty.³

Bohumír Dejmek při vytváření frekvenčního slovníku mluvené češtiny definoval mluvenost takto: „Běžně mluveným jazykem rozumíme soubor mluvených prostředků (spisovných i nespisovných) fungujících ve vymezené komunikační sféře, v projevech vyznačujících se *komplexní mluveností*. Je poznamenán výběrem slovní zásoby a specifickými stylistickými prostředky (využití slov parazitních,

¹ Daneš, F.: Perspektivy češtiny. In: *K diferenciaci současného mluveného jazyka*. Ostrava 1995, s. 6.

² Chloupek, J.: Sjednocující a rozrůzňující faktory v mluvené komunikaci. In: *K diferenciaci současného mluveného jazyka*. Ostrava 1995, s. 37.

³ Kraus, J.: Perspektivy češtiny. In: *K diferenciaci současného mluveného jazyka*. Ostrava 1995, s. 12–18.

kontaktoých), častým opakováním slov, různými nepravidelnostmi větné stavby, častými přerušováními, nedokončením projevu.“⁴

V této práci nám jde primárně o zkoumání spontánního, dialogického a neformálního mluveného jazyka, tedy se můžeme přiklonit k Danešově definici *jazyka běžně mluveného*. V žádném případě tak nebude předmětem našeho zájmu mluvený jazyk tak, jak byl chápán například v některých starších frekvenčních slovnících (viz níže).

2.2 Psanost vs. mluvenost

Stejně jako není snadné přesně definovat pojmy *psanost* a *mluvenost*, není ani jednoduché jednoznačně vymežit vztah obou forem realizace jazyka. Světlá Čmejková mluví o třech základních možných vztazích psanosti a mluvenosti:

- „dvě rozdílné kulturní techniky a sociální aktivity (ústnost a gramotnost);
- dva autonomní jazyky;
- dva různé existenční mody.“⁵

Vztah psanosti a mluvenosti asi nejlépe vystihneme, když si uvědomíme, že mluvené texty zde existovaly daleko dříve, psané texty přišly mnohem později, a to jako záznam mluvených textů. Máme tedy za to, že mluvená forma je primární a psaná sekundární. Toto pojetí dnes uznává většina lingvistů, přičemž je třeba připomenout, že v dřívějších dobách směřovala pozornost jazykovědců výhradně k formě psané, mluvená byla buď na okraji, nebo zcela mimo zájem lingvistiky. Obrat můžeme sledovat již od 1. pol. 20. stol., kdy Ferdinand de Saussure tvrdil, že jediný důvod, pro který psaní existuje, je ten, aby reprezentovalo řeč. Podobná tvrzení následně zastávali také ostatní významní lingvisté jako třeba Leonard Bloomfield nebo Edward Sapir. Na druhou stranu je třeba říct, že byly i jiné názory, avšak ve výrazné

⁴ Dejmek, B.: Frekvence slov a tvarů substantiv a sloves v běžně mluveném jazyce. *Slovo a slovesnost*, 63, 2002, s. 15.

⁵ Čmejková, S.: Mluvenost a psanost. In: *Přednášky z doktorandských dnů*. Praha 2006, s. 56.

menšině, nejvýraznějším odpůrcem tzv. fonocentrického pohledu byl francouzský lingvista Jacques Derrida. Dalším významným mezníkem byl pak nástup počítačových technologií, využívaným hlavně v korpusové lingvistice.

Fakt, že mluvenou formu považujeme za primární, neznamena, že by měla být psaná formě věnována menší pozornost, jen poukazujeme na skutečnost, že psaná forma vznikla jako záznam mluvených projevů. Během vývoje lidské společnosti však psanost dosáhla takové důležitosti, že právě psaná forma značně ovlivňuje mluvenou. Odpovědět tak na otázku, která forma více ovlivňuje tu druhou, je velmi obtížné, ba nemožné.

2.3 Podstata diference psaných a mluvených textů

Jestliže jsme vymezili psanost a mluvenost a popsali vztahy mezi nimi, nyní se budeme zabývat tím, které faktory určují rozdíly mezi oběma formami realizace jazyka.

Primárně se považuje za základní a někdy i jediný prvek determinující odlišnosti textů mluvených od psaných samotná zvuková realizace. Tento faktor považujeme také za dominantní, ale je třeba upozornit na fakt, že samotná zvuková realizace není jediným faktorem určujícím rozdíly psaných a mluvených textů.

Jestliže přistoupíme na výše vymezené chápání mluveného jazyka, tedy jako spontánní neformální projev, zpravidla dialogický, je zřejmé, že kromě samotné zvukové realizace zde vstupují do hry také jiné faktory, které již nejsou součástí obecné lingvistiky a jejích tradičních disciplín, ale spadají spíše do nových pomezích disciplín jako je sociolingvistika, psycholingvistika či pragmalingvistika. Tento fakt tak značně komplikuje jednostranné pojmání mluvenosti jakožto formy jazyka determinované výhradně svou zvukovou realizací.

Za základní opozitní charakteristiky obou forem jsou považovány připravenost psaných a naproti tomu spontánnost mluvených textů. S tímto tvrzením lze v zásadě souhlasit, ale je třeba upozornit, že situace se začíná v současné době značně komplikovat z důvodů již zmíněné interference psanosti a mluvenosti u mnoha textů, které se dříve vyskytovaly v jen omezené míře nebo vůbec. Například s nárůstem

tzv. elektronické komunikace (e-mail, sms, chat) do psaných textů výrazně proniká spontánnost; připravenost (s tím i např. formálnost, oficiálnost, plánovitost) tak v těchto psaných textech přestává být dominantní. K podobné situaci dochází u některých mluvených textů, což souvisí s již zmíněným nárůstem významu schopnosti kultivovaně a účinně se vyjadřovat, tento trend bývá označován jako renesance rétoriky. Např. v mluvených textech na poli odborném, obchodním nebo diplomatickém nepochybně dochází k výraznému oslabování spontánnosti a nárůstu připravenosti. Je tak patrné, že v mnoha ohledech dochází k prolínání různých charakteristik, které jsou tradičně přisuzovány psanosti či mluvenosti.⁶

Přestože hranice psaných a mluvených textů jsou neostré, a rozdíl tak nejsou příliš vyhraněné, neznamená to, že by nebyly vůbec či že by nemělo smysl se jimi zabývat. Naopak považujeme právě v této době za důležité věnovat oněm rozdílům maximální pozornost, neboť jazykověda má být vědou především deskriptivní a teprve druhotně preskriptivní. Pouze analyzováním reálných textů můžeme pravdivě popisovat jak stav současného jazyka, tak jeho vývoj.

⁶ Krčmová, M.: K pramenům specifika mluveného projevu (mluvenost či spontánnost?). In: *K diferenciaci současného mluveného jazyka*. Ostrava 1995, s. 26–32.

3 METODOLOGIE LEXIKÁLNÍ ANALÝZY

Při vytváření jazykových korpusů a zásad jejich zpracování jsme vycházeli především ze dvou zdrojů. Prvním je publikace Marie Těšitelové *Kvantitativní lingvistika*, ve které jsme hledali poučení zejména při sestavování textů k analýze, ale také zásady statistického zpracování lingvistických dat. Druhým důležitým zdrojem byl *Frekvenční slovník češtiny* (FSČ), který nám byl oporou především při lemmatizaci slovníku. Kromě těchto dvou publikací jsme při sestavování frekvenčních slovníků přihlíželi také k *Frekvenčnímu slovníku mluvené češtiny* (FSMČ).

3.1 Výběr a rozsah materiálu (korpusu)

Prvním krokem při vytváření jakékoliv kvantitativní analýzy je výběr vhodného korpusu textů, který je základem relevance později získaných výsledků. Pro jakoukoliv statistiku je základním parametrem rozsah, kde vždy platí, čím větší, tím větší relevance získaných dat. Druhým neméně důležitým parametrem při vytváření jazykového korpusu je kvalitativní selekce. Výběr materiálu tak vždy provádíme se zřetelem ke dvěma základním hlediskům – kvantitativnímu a kvalitativnímu.

V našem případě budeme vytvářet dva samostatné korpusy – psaný a mluvený, které budou separátně zkoumány a podrobovány lexikální analýze. Pro relevanci výsledků a jejich následnou komparaci je především důležité, aby oba soubory textů byly:

- a) stejné co do rozsahu,
- b) zpracovávány podle stejných pravidel a zásad (viz dále).

3.1.1 Kvantitativní hledisko při vytváření korpusu

Stanovení rozsahu textů je základním problémem každé lingvistické analýzy opírající se o statistiku. Jak již bylo uvedeno výše, ideální pro statistiku je zpracování veškerých existujících textů daného jazyka, což je samozřejmě nereálné zejména kvůli technickým limitům. Proto je důležité přistupovat k otázce rozsahu zkoumaného materiálu velmi zodpovědně, neboť relevance veškerých výstupů z něj získaných je jím přímo ovlivněna.

Vzhledem k omezujícím parametrům – ať už technickým nebo časovým – a se zřetelem k základním cílům této práce jsme stanovili pro oba korpusy rozsah 20 000 slov.⁷

3.1.2 Kvalitativní hledisko při vytváření korpusu

Kvalitativní hledisko spočívá v otázce, jaké zásady budou užity při výběru textů, tedy které texty do korpusu zařadit, a které ne. Určit správná pravidla pro kvalitativní výběr je velmi obtížné, protože stejně jako kvantitativní rozsah výrazně určuje relevanci pozdějších výsledků. Zásadní pro kvalitativní výběr je především účel, kterému má lingvistická analýza sloužit. Těšitelová pro kvalitativní výběr rozlišuje šest kritérií:

1. „jazyková,
2. psychologická,
3. sociologická,
4. tematická,
5. sémiotická,
6. jiná.“⁸

⁷ Myšleno tzv. textových slov, tedy písmene nebo řady písmen mezi mezerami.

⁸ Těšitelová, M.: *Kvantitativní lingvistika*. Praha 1987, s. 19.

3.1.2.1 Psaný korpus

Pro naši potřebu jsme zvolili pro psaný korpus texty z jednoho vydání deníku Mladá fronta Dnes, konkrétně ze dne 1. prosince 2009. Tento zdroj jsme vybrali především z důvodu, že se jedná o největší české seriózní noviny, které tak představují reprezentativní vzorek současného psaného jazyka. Použity byly texty jak zpravodajské, tak publicistické (v užším smyslu). Publicistický funkční styl shledáváme nejvhodnějším pro naše účely, neboť funkční styl umělecký, administrativní nebo odborný rozhodně dostatečně nereprezentují současný stav většinového úzu a také nejsou příliš vhodné pro naši komparaci s mluvenými texty.

Texty byly excerpovány z téměř všech rubrik (domácí a zahraniční zpravodajství, publicistika, ekonomika, kultura, regionální zpravodajství – v našem případě středomoravská mutace regionální přílohy), nezařadili jsme texty ze sportovní rubriky, které mají značně specifický charakter a zdají se tak být pro naše účely nejméně vhodné. Dále jsme nezařazovali jakékoliv texty inzertního charakteru, tzn. i články, jež se svým formátem blíží redakčním článkům, ale jsou pouze komerčním sdělením. Poslední kategorií textů, které jsme nezařadili do našeho výběru, jsou všechny formy tzv. infografiky, kam patří nejrůznější výčty, grafy, tabulky apod., které slouží jako opora či dodatečná informace k článkům, tyto texty mají své zvláštní specifické charakteristiky, které jsou pro naše účely naprosto nevhodné.

Důležité je také uvést charakter podavatelů vybraných textů, kterými jsou téměř výhradně profesionální redaktoři, u nichž se počítá s vysokou jazykovou kompetencí, tzn. bezvadnou znalostí a aplikací stylistických, syntaktických a ortografických norem. Na druhou stranu je kreativita novinářů značně omezována interními metodickými pokyny, které mají za cíl především maximální míru srozumitelnosti pro všechny čtenáře, jejichž jazyková kompetence je různá, zpravidla však vždy mnohem nižší než redaktorova. Proto institucionální podavatel (mediální instituce) emituje taková nařízení, jež mají za cíl minimalizovat náročnost textů ve všech jazykových plánech vzhledem k usnadnění percepce čtenáři. Tím se současný publicistický styl diverzifikuje od stylu odborného, uměleckého a administrativního.

Novinové články tak mají, jak již bylo uvedeno výše, nejbližší k současnému úzu a zdají se být nejhodnějším materiálem pro naši analýzu současné psané češtiny.

3.1.2.2 Mluvený korpus

Při výběru druhého korpusu, tedy mluveného, byly možnosti mnohem omezenější než v případě psaného. Důvody jsou zřejmé, mluvené texty zpravidla nejsou nikde zaznamenávány, pokud ano, jedná se zpravidla o projevy předem připravené, tedy nejde o čistě mluvené texty. Pro potřeby lingvistů tak vzniklo několik publikací, které obsahují transkripty autentických mluvených projevů. V našem případě jsme použili práci Ireny Bogoczové,⁹ která obsahuje mluvené texty z různých oblastí České republiky, přičemž mluvčí mají různé vzdělání, sociální postavení, věk apod. Je důležité uvést, že se jedná o texty zpravidla dialogické, neboť monologické projevy nejsou reprezentativním vzorkem většinového úzu a spadají spíše na periferii mluvenosti, protože jsou často předem připravené.

Konfrontaci těchto dvou materiálů považujeme za vhodnou vzhledem k cílům této práce. Oba korpusy představují aktuální jazyk, který reprezentuje většinový úzus jak psané, tak mluvené formy. Jejich komparace by tak měla odhalit relativně relevantní reprezentativní výsledky výstupních dat, které by měly ukázat rozdíly současné psané a mluvené češtiny ve sledovaných parametrech.

3.2 Lexikální jednotka

Při zpracovávání lingvistických dat je po vhodném výběru jazykového materiálu asi nejdůležitější, zvláště pak v lexikální statistice, definování jednotky souboru. Tato problematika se v různých jazycích značně liší, zejména vzhledem k jejich

⁹ Bogoczová, I. a kol.: *Tváře češtiny*. Ostrava 2000.

morfologické typologii. Zvlášť u flektivních jazyků je pro definici jednotky souboru důležité rozlišení pojmů *slovo* a *lexém*.

Slovo zde chápeme v podstatě jako tvar slova (textové slovo), tedy jako písmeno či řadu písmen mezi mezerami. *Lexém* chápeme jako lexikální jednotku, která je dána základním tvarem slova, tedy slovníkovým heslem (lemmatem), jež reprezentuje všechny různé realizace (tvary) jednoho lexému, v případě jmen je to nom. sg., u sloves infinitiv.

Pro další zpracování lingvistických dat a výpočet žádaných charakteristik (rozsah, rozptýlení a koncentrace lexika) bylo třeba vytvořit z psaného a mluveného korpusu seznamy lemmat, které jsou řazeny sestupně dle jejich absolutní frekvence (četnosti jejich výskytu v daném souboru textů). Při zpracování těchto seznamů jsme dodržovali základní zásadu kvantitativní lingvistiky, kdy každá jednotka se někde započítává a žádná jednotka se nepočítá dvakrát. Dále byla vzhledem k následné komparaci obou seznamů a dalších dat z nich vypočítaných důsledně dodržována zásada aplikace stejných pravidel při zpracovávání obou výchozích textů. Tato pravidla definujeme v následujících kapitolách.

3.2.1 Negace

V záporné podobě uvádíme pouze slova, která nemají svoji kladnou podobu, nebo se negativní podoby užívá častěji, popř. se jí užívá v jiném významu. Žádné slovo jednoho významu se tak nesmí v seznamu objevit v záporné i kladné podobě, musí být stanovena vždy jedna základní podoba, do níž se započítávají všechny tvary.

3.2.2 Stupňování adjektiv a adverbíí

V češtině se při lemmatizaci vyskytuje problém stupňování adjektiv a adverbíí. V této práci, stejně jako ve FSČ, nepovažujeme komparativ a superlativ za zvláštní jednotky, a řadíme je tak v seznamu pod tvar pozitivu.

3.2.3 Zájmena

Při zpracování zájmen jsme postupovali analogicky k ostatním zásadám zpracování frekvenčního slovníku a definovali základní jednotky, pod něž se započítávají jejich ostatní tvary. U osobních zájmen považujeme za různé jednotky pouze *já, ty, on, my, vy, oni*, pod které jsou zahrnuty ostatní tvary. U přivlastňovacích zájmen postupujeme analogicky a za různé jednotky považujeme *můj, tvůj, jeho, náš, váš, jejich*, ostatní tvary jsou opět řazeny pod ně.

Ukazovací zájmeno *ten* ve slovníku reprezentuje všechny jeho modifikace jako *ta, to, ti, ty, ta, toho, těch* apod. Je však nutné zde uvést, že pod zájmeno *ten* neřadíme zájmena *tenhle* a *tenhleten*, která považujeme stejně jako FSČ za různé jednotky.

Zájmena *se, si* započítáváme jako samostatnou jednotku pouze pokud měla v textu funkci předmětu, tzn. šlo je nahradit tvary *sebe, sobě*. V ostatních případech, zvláště u reflexiv tantum, se zájmena *se, si* nezapočítávají jako samostatné jednotky, nýbrž je řadíme pod sloveso.

3.2.4 Sloveso být

Při statistickém zpracování lingvistických dat v češtině je značně problematické zařazování slovesa *být*, přičemž nejobtížnější je rozlišení a započítávání pasiva složeného slovesného predikátu, kde je vyjádřen výsledek děje, a slovesně-jmenného predikátu, který vyjadřuje stav nebo vlastnost. Pasivní tvary (např. stroj byl vypnut) považujeme za jednu lexikální jednotku, zatímco slovesně-jmenné predikáty (např. stroj byl vypnutý) považujeme za dvě lexikální jednotky.

3.2.5 Předložky

Předložky započítáváme jako samostatné jednotky.

3.3 Úskalí statistiky

Statistika je metoda aplikované matematiky, která se uplatňuje snad ve všech oborech. Vedle svých nesporných výhod má však statistika své limity a úskalí, o nichž je třeba se zmínit.

Lexikální a vůbec celá jazyková statistika naráží na základní problém nemožnosti obsažení všech textů, tento problém je také nejzávažnějším faktorem snižujícím relevanci výstupních dat našeho nejstaršího frekvenčního slovníku *Frekvence slov, slovních druhů a tvarů v českém jazyce* (FSSDTČJ). Závažnost tohoto omezujícího faktoru značně snížil vývoj počítačových technologií a vytvoření speciálních softwarových nástrojů pro zpracování lingvistických dat. V případě FSC je problematika rozsahu téměř, nikoliv zcela, vyřešena, neboť se analyzoval jazykový materiál s počtem 100 milionů slov, v případě FSMČ pak 780 tisíc slov. Vzhledem k rychlému vývoji počítačové lingvistiky bude možné zpracovávat stále obsáhlejší materiály. Tradiční problematika jakékoliv statistiky se tak v případě jazykové zdá být takřka vyřešena. Problematictější a počítačově těžko řešitelný je aspekt kvalitativního výběru textů, o kterém jsme se již zmínili. Např. pro FSC bylo složení materiálu stanoveno dle funkčních stylů tak, že 15 % tvoří umělecký, 25 % odborný a 60 % publicistický. Přestože byl tento poměr stanoven na základě několika průzkumů, těžko může někdo tvrdit, že jde o stoprocentně reprezentativní vzorek současné psané češtiny, navíc výběr beletristických textů má mnohá specifika, která dále oslabují reprezentativnost materiálu.

Statistika tak nemůže postihnout celou realitu, ale snaží se na základě analýzy vhodně vybraného materiálu realitě co nejvíce přiblížit. Je však obtížné stanovit, do jaké míry zvolená pravidla výběru a zpracování materiálu odpovídají skutečné jazykové situaci. V této práci se samozřejmě nemůžeme srovnávat s obsáhlostí jak FSC, tak FSMČ, ale vzhledem k našim cílům a možnostem považujeme náš výběr za dostatečný a jako ověření správnosti naší volby můžeme použít zmíněné dva slovníky jako vhodný materiál ke komparaci výstupních dat.

3.4 Některé práce zabývající se frekvenční analýzou češtiny

Je zřejmé, že psané češtině se lingvisté věnují mnohem déle a důsledněji, teprve v poslední době se zvyšuje pozornost mluvené podobě jazyka. Proto i práce zpracovávající statisticky psané texty nutně počtem převyšují práce zpracovávající mluvenou češtinu, tento fakt je pak ještě umocněn technickými omezeními, kdy psané texty jsou mnohem dosažitelnější než mluvené, které se musí složitě získávat nahráváním a jejich následným přepisováním.

3.4.1 Psaná čeština

V statistickém zpracování dat je důležitým mezníkem náš vůbec první frekvenční slovník *Frekvence slov, slovních druhů a tvarů v českém jazyce* (FSSDTČJ), který se stal na dlouhá léta základní prací svého druhu. Tento slovník byl zpracován na základě materiálu ze 75 publikací o rozsahu 1,6 milionu slov. Tato práce již obsahovala oba základní slovníky – řazené podle frekvence a abecedy. Kromě toho tato práce obsahovala také další údaje, např. frekvenci pádů u substantiv nebo slovesných tvarů.

Další práce tohoto druhu přišla v roce 1983, jedná se o *Frekvenční slovník češtiny věcného stylu*, který byl sestaven ještě na omezenějším korpusu než FSSDTČJ.

Teprve v roce 2005 byl FSSDTČJ překonán *Frekvenčním slovníkem češtiny* (FSC), který vznikl jako kolektivní dílo Ústavu Českého národního korpusu FF UK v Praze. Tento slovník zpracovává materiál o rozsahu sta milionů slov a vychází z korpusu SYN2000. Kromě tradičně udávané absolutní frekvence, tento slovník udává také hodnotu tzv. průměrné redukované frekvence, která se zaměřuje i na rozložení výskytu slova v celém korpusu, to znamená, že zohledňuje i počet pramenů, v nichž se slovo vyskytuje.

3.4.2 Mluvená čeština

V případě mluvené češtiny můžeme uvést v podstatě jedinou publikaci – *Frekvenční slovník mluvené češtiny*, který je vůbec první prací svého druhu, tedy zpracovávající mluvený jazyk. Výchozím materiálem je Pražský mluvený korpus, nejvíce tak zde je zastoupena obecná čeština, která je nejužívanější varietou mluvené češtiny v celé České republice.

V kontextu mluvené češtiny je třeba zmínit fakt, že už autoři FSSDTČJ do zkoumaného materiálu zařadili část mluvených textů, které však zahrnovaly pouze dramata a oficiální projevy tehdejších představitelů moci. Proto jako první frekvenční slovník zpracovávající skutečně mluvenou češtinu chápeme až FSMČ.

Pro úplnost můžeme ještě zmínit *Frekvenční slovník češtiny věcného stylu* (FSVS) a publikaci Marie Těšitelové *Psaná a mluvená odborná čeština z kvantitativního hlediska*, která analyzuje data FSVS. Tato práce však v rámci mluvených textů analyzuje pouze texty patřící do odborného funkčního stylu (jde zejména o přednášky, referáty apod.), které byly mluvčími předem připraveny, tudíž nešlo o analýzu běžně mluvené češtiny. Přesto již tato práce odhalila na základě komparace statistických dat jisté odchylky mezi psanými a mluvenými (i když předem připravenými) texty. Výsledné hodnoty budeme srovnávat s našimi výstupními hodnotami (viz níže).

4 ROZSAH LEXIKA

4.1 Definice

Rozsah lexika, nazývaný též bohatství slovníku, je kvantitativní charakteristika, která vyjadřuje vztah mezi délkou textu a jeho slovníkem, kde délkou textu je myšlen počet slov (tedy písmenem či řadou písmen mezi mezerami, v našem případě tedy 20 000 slov v každém korpusu, viz výše), slovníkem je pak myšlen počet lexikálních jednotek (viz výše).

Základní rovnici pro výpočet bohatství slovníku stanovil francouzský lingvista P. Guiraud:

$$R = V / \sqrt{N}$$

R – bohatství slovníku (rozsah lexika)

V – slovník

N – délka textu

Tato základní rovnice pro výpočet bohatství slovníku Těšitelová modifikuje: „[...] je třeba – pro češtinu – počítat jen s 80 % textu, tj. pouze se slovy plnovýznamovými (substantiva, adjektiva, zájmena, číslovky, slovesa a adverbia), popř. se 70 % textu, pokud za slova plnovýznamová pokládáme jen substantiva, adjektiva, slovesa a adverbia.“¹⁰

Takto modifikované rovnice pak vypadají následovně:

¹⁰ Těšitelová, M.: *Kvantitativní lingvistika*. Praha 1987, s. 71.

$$R = 100 \times (V / {}^8/_{10} \times N)$$

$$R = 100 \times (V / {}^7/_{10} \times N)$$

Stejně jako Těšitelová budeme primárně pracovat s první rovnicí, která počítá s 80 % plnovýznamových slov.

4.2 Výpočet

Výpočet rozsahu lexika pro psaný korpus:

$$R_p = 100 \times (V / {}^8/_{10} \times N)$$

$$R_p = 100 \times (5\,201 / {}^8/_{10} \times 20\,000)$$

$$R_p = 32,5$$

Výpočet rozsahu lexika pro mluvený korpus:

$$R_m = 100 \times (V / {}^8/_{10} \times N)$$

$$R_m = 100 \times (3\,062 / {}^8/_{10} \times 20\,000)$$

$$R_m = 19,1$$

4.3 Analýza

Rozsah slovníku patří mezi elementární charakteristiky získané aplikací lexikální analýzy, výstupní hodnoty nám ukazují v podstatě index opakování slov. Získané výsledky tak např. ukazují na rozdílnost této charakteristiky v závislosti na funkčním stylu analyzovaného textu, nám však půjde o srovnání výsledných hodnot v závislosti na psané a mluvené formě komunikátu.

Rozdíl získaných hodnot je zřejmý, psané texty (v našem případě spadající do publicistického funkčního stylu) vykazují mnohem větší bohatství slovníku než je tomu v případě textů mluvených. Faktorů, které determinovaly tyto hodnoty v naší analýze je hned několik.

4.3.1 Připravenost-nepřipravenost

Prvním důležitým faktorem je opozice připravenosti-spontánnosti produkce komunikátu. Zatímco psané texty byly vždy důkladně připraveny a zpravidla prošly několikerou úpravou, texty mluvené byly takřka vždy zcela spontánní, a tudíž nepřipravené. Podavatel tak v druhém případě nutně užívá menšího slovníku, neboť je limitován, zvláště při dialogu, časem na formulování výpovědi, což způsobuje méně promyšlené realizace. Produktor textu při časovém tlaku totiž užívá jen menší část svého idiolektu, která stojí v jeho centru, dochází tak k nadměrnému užívání stejných výrazů. Index opakování slov tak dosahuje vysokých hodnot a dochází tak nutně k malému bohatství slovníku.

Problematika připravenosti úzce souvisí, jak již bylo naznačeno, s množstvím času, které má podavatel textu k dispozici pro realizaci výpovědi. Považujeme však za vhodné zde upozornit, že čas není jediným, přestože dominantním, činitelem hrajícím roli v připravenosti. Při mluvené výpovědi také zpravidla není možné získat jiné podpůrné informace než ty, které máme v mozku, při produkci psaných textů je však často užíváno různých zdrojů, které následně obohatí výsledný text. V našem případě, kdy psaný korpus reprezentují výhradně texty publicistického funkčního stylu, je tento činitel velmi významný, neboť redaktor takřka vždy tvoří texty na základě textů jiných, někdy dochází dokonce k pouhé modifikaci či sloučení jiných textů.

4.3.2 Charakter podavatele

Dalším významným faktorem ovlivňujícím rozsah slovníku je zvlášť v našem případě charakter podavatele textu, nejrůznější individuální rysy každého jednotlivce ovlivňující jeho vyjadřování zde opomeneme a pozornost soustředíme na jazykovou kompetenci, především ve smyslu profesionality podavatele. V našich dvou korpusech tak z tohoto hlediska dochází k markantnímu kontrastu, neboť v případě mluvených textů je zpravidla podavatelem průměrně vzdělaný člověk bez zvláštních znalostí českého jazyka a komunikačních dovedností, v případě psaných textů se takřka vždy jedná o profesionální redaktory, kteří disponují nadprůměrnou jazykovou kompetencí, získanou jednak svým vzděláním, jednak komunikační zručností. Žurnalista tak šíří svého idiolektu zpravidla vždy výrazně převyšuje průměrného člověka, to se nutně musí projevit v bohatství slovníku jím realizovaných textů.

V případě publicistického stylu je charakter podavatele velmi komplikovaný, neboť zde spíše než o novináři je vhodnější mluvit o institucionálním podavateli. Tento výraz totiž mnohem lépe vystihuje charakter subjektu tvořícího výsledný text, neboť téměř vždy (v případě článků MF Dnes můžeme říct zcela vždy) se na výsledné podobě textu podílí hned několik lidí. Pro přehlednost jsme vytvořili následující schéma:

událost	→	podpůrné zdroje informací	→	redaktor	→	korektor	→	editor
tisková zpráva		jiná média						
komerční sdělení		vyjádření účastníků události						
tisková konference		vyjádření odborníků						
skutečná událost		vyjádření tiskových mluvčích						
fiktivní událost		literatura						
událost v jiném médiu		internetové zdroje						
redaktorova intence		rešerše						

Toto schéma je obecné, konkrétní realizace může být různě modifikována, ve většině případů však těmito procesy prochází vytváření většiny článků, zvláště pak v případě velkých vydavatelských domů, kam patří i námi sledovaná MF Dnes.

Ze schématu je patrné, že na vytváření článku se nepodílí jen jeden profesionální žurnalista, ale také další subjekty. Důležitým specifikem publicistických textů je také vkládání různých vyjádření v podobě přímých citací, kdy do textu vstupuje mnohdy i několik dalších subjektů v rámci jednoho článku. Všechny tyto faktory výrazně působí na zvýšenou hodnotu rozsahu slovníku analyzovaného textu.

4.3.3 Téma textů

Nepochybně výrazným činitelem ovlivňujícím bohatství slovníku je téma textů. Již v úvodní části této práce v kapitole věnované definici mluvenosti jsme vymezili naše pojetí analyzovaných mluvených textů, jejichž obsah se zpravidla vztahuje k běžným společenským jevům. Přestože jsou psané publicistické texty určeny pro nejširší vrstvy obyvatel a věnují se tak aktuálním společenským událostem z různých oborů lidské činnosti, je zřejmé, že tematika je poněkud rozmanitější, než je tomu v případě běžné komunikace většinové populace, navíc veškerá témata jsou popisována více do hloubky. Tento fakt tak zvyšuje rozsah lexika ve prospěch psaných komunikátů.

4.3.4 Oficiálnost-soukromost

Obecně platí, že každý psaný text je potenciálně veřejný, kdežto u mluvených textů se naopak předpokládá soukromý charakter. Tato skutečnost ovlivňuje také rozsah slovníku, protože podavatel oficiálního sdělení se vždy snaží vyjádřit se více obsáhle.

4.3.5 Explicitnost a určitost

Při produkci psaného projevu se podavatel vždy vyjadřuje více explicitně a určitě než u mluveného. Důvodů je několik, mezi nejdůležitější patří tyto:

- Pisatel má zpravidla mnohem větší prostor pro vyjádření.
- Možnost kontroly a zpětné opravy již napsaného.
- Pisatel vždy klade důraz na jednoznačnost a úplnost vyjádření, všechny potenciální nejasnosti totiž mohou mít negativní vliv na percepci textu.
- V případě mluveného komunikátu má podavatel zpravidla vždy zpětnou odezvu publika, což mu umožňuje pružně reagovat a přizpůsobovat následující výpověď.
- V mluveném projevu podavatel užívá vedle znakového systému přirozeného jazyka také další výrazové prostředky, kam patří jednak parajazykový systém neverbálních vokálních prozodických suprasegmentálních prostředků (větná intonace, frázování řeči, větný důraz, mluvní tempo, zabarvení hlasu) a jednak systém neverbálních nevokálních prostředků (proxemika, haptika, posturika, kinezika, mimika, pohledy). Tím, že mluvčí omezuje postavení přirozeného jazyka a zapojuje další systémy, dochází v lexikální rovině k menší explicitnosti a určitosti.

Všechny uvedené činitele tak mají výrazný vliv na rozsah slovníku, protože méně explicitně a určitě vytvořený text má jednoznačně chudší lexikum.

4.3.6 Abstrakce

Psané projevy mají zpravidla vždy větší míru abstrakce v pojmenování skutečnosti. Tento fakt úzce souvisí s předešlými výše uvedenými faktory ovlivňujícími rozsah lexika. Zejména tento jev způsobuje větší časový prostor pro formulování myšlenek, také psané texty jsou obecně delší než mluvené, roli zde hraje také zmíněná oficiálnost a explicitnost. Je zřejmé, že většina textů vyjadřujících složitější myšlenky má psanou podobu, nejvíce patrný je tento fakt u odborného funkčního stylu, jehož texty jsou z drtivé většiny psané, ostatní mluvené texty jako

přednášky nebo referáty navíc nemůžeme hodnotit jako jednoznačně mluvené, protože jde o důkladně připravené komunikáty, které můžeme spíše než jako mluvené označit jako čtené psané texty, které však svoji aktuální zvukovou realizací také mají prvky mluveného projevu, nejedná se však v rozhodně, jak již bylo uvedeno, o typické mluvené texty.

4.3.7 Organizovanost

Míra organizace textu bezprostředně souvisí i s bohatstvím lexika, neboť vysoce organizovaný text nutně potřebuje bohaté vyjadřovací prostředky. Z hlediska organizovanosti jsou výše jednoznačně psané texty, jejichž charakter to vzhledem k funkci vyžaduje a zároveň výše uvedenými faktory umožňuje.

Rozsah lexika je evidentně mnohem vyšší u psaných textů než mluvených, to jsme dokázali jak samotným výpočtem, tak následnými rozbory jednotlivých určujících faktorů.

5 ROZPTÝLENÍ LEXIKA

5.1 Definice

Rozptýlení lexika (slovníku) vyjadřuje poměr počtu plnovýznamových slov s frekvencí 1–10 k slovníku. Základní rovnice pro výpočet:

$$D = 100 \times (V_1^{10} / V)$$

D – rozptýlení slovníku

V_1 – počet plnovýznamových slov s frekvencí 1–10

V – slovník

5.2 Výpočet

Výpočet rozptýlení lexika pro psaný korpus:

$$D_p = 100 \times (V_1^{10} / V)$$

$$D_p = 100 \times (4\,959 / 5\,201)$$

$$D_p = 95,3$$

Výpočet rozptýlení lexika pro mluvený korpus:

$$D_m = 100 \times (V_1^{10} / V)$$

$$D_m = 100 \times (2\,852 / 3\,062)$$

$$D_m = 93,1$$

5.3 Analýza

Z námi zjištěných hodnot vyplývá, že charakteristika rozptýlení lexika není ve vztahu difference psané a mluvené češtině relevantní, rozdíl hodnot D_p a D_m je zanedbatelný, vzhledem ke všem faktorům ovlivňujícím statistiku jej můžeme považovat dokonce za nulový.

Rozptýlení lexika, tedy poměr počtu plnovýznamových slov s frekvencí 1–10 k slovníku, tak s největší pravděpodobností je spíše charakteristikou jazyka jako konkrétního přirozeného systému znaků sloužícího ke komunikaci jednoho jazykového společenství. Nabízí se tak otázka, do jaké míry se hodnota této charakteristiky liší v různých jazycích, zejména pak ve srovnání jazyků různé typologie. Takový výzkum si však svým rozsahem žádá samostatnou podrobnou práci.

6 KONCENTRACE LEXIKA

6.1 Definice

Koncentrace lexika (slovníku) vyjadřuje poměr délky textu odpovídající prvním 10 nejfrekventovanějším slovům k celkové délce textu. Základní rovnice pro výpočet:

$$K = 100 \times (N_1^{10} / N)$$

K – koncentrace slovníku

N_1 – délka textu odpovídající prvním 10 nejfrekventovanějším slovům

N – délka textu

6.2 Výpočet

Výpočet koncentrace lexika pro psaný korpus:

$$K_p = 100 \times (N_1^{10} / N)$$

$$K_p = 100 \times (3\,007 / 20\,000)$$

$$K_p = 15,04$$

Výpočet koncentrace lexika pro mluvený korpus:

$$K_m = 100 \times (N_1^{10} / N)$$

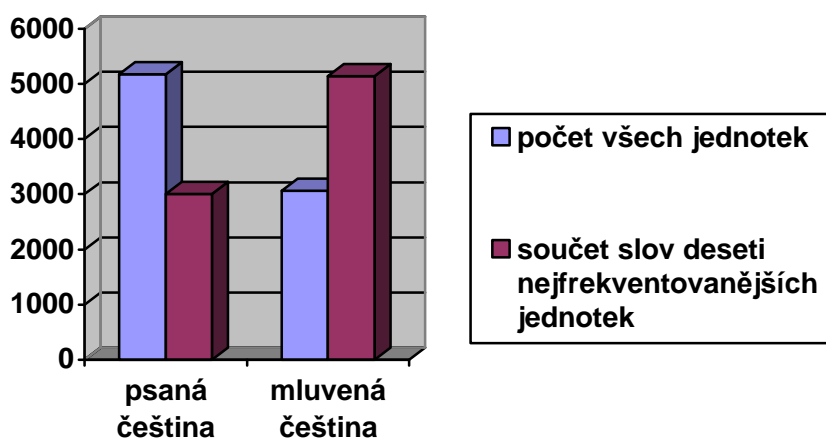
$$K_m = 100 \times (5\,144 / 20\,000)$$

$$K_m = 25,72$$

6.3 Analýza

Rozdíl výsledných hodnot je značný, mluvená čeština dosahuje 171 % hodnoty koncentrace lexika češtiny psané. Poměr délky textu odpovídající prvním 10 nejfrekventovanějším slovům k celkové délce textu tak výrazně determinuje diference mezi psanou a mluvenou češtinou.

Základním jevem určujícím tuto rozdílnost je fakt, že psaná čeština má mnohem větší množinu jednotek než mluvená. Zatímco psaná čeština v našem slovníku obsahuje 5 179 jednotek, v případě mluvené je to 3 062, psaná čeština má tak mnohem větší množinu slov, což bezprostředně souvisí s rozsahem slov (viz výše). Jestliže množina lexikálních jednotek mluvené češtiny dosahuje 59 %, tedy téměř poloviny, psané, je nutné k vytvoření stejně dlouhého textu mnohem více užívat stejných jednotek. Koncentrace mluvené češtiny je tak mnohem vyšší. Pro názornost jsme vytvořili následující graf:



Z grafu je patrné, že mezi celkovým počtem jednotek ve slovníku a součtem slov deseti nejfrekventovanějších jednotek je vztah nepřímé úměry – čím větší je počet jednotek ve slovníku, tím menší je počet deset i nejfrekventovanějších slov. Zatímco v případě psané češtiny mají tyto dva parametry hodnotu 5 179 a 3 007, v případě mluvené češtiny jsou tyto hodnoty téměř zcela opačné – 3 062 a 5 144. Na základě těchto výsledků pak můžeme stanovit obecnou rovnici postihující tento vztah:

$$\Sigma_1 \times N_1^{10} = \Sigma_2 \times N_2^{10}$$

N_1^{10} – délka textu odpovídající prvním 10 nejfrekventovanějším jednotkám prvního korpusu

N_2^{10} – délka textu odpovídající prvním 10 nejfrekventovanějším jednotkám druhého korpusu

Σ_1 – suma jednotek prvního korpusu

Σ_2 – suma jednotek prvního korpusu

Jestliže dosadíme do rovnice naše hodnoty, výpočet vypadá takto:

$$\Sigma_1 \times N_1^{10} = \Sigma_2 \times N_2^{10}$$

$$5\,179 \times 3062 = 3007 \times 5144$$

$$\mathbf{15\,858\,098 = 15\,468\,008}$$

Vedle této negativní korelace jsme také zjistili, že poměr sumy jednotek v textu násobený hodnotou jedné setiny jeho koncentrace je konstantní, tento vztah můžeme pak definovat rovnicí:

$$\Sigma \times \frac{1}{100} K = k$$

Σ – suma všech jednotek v korpusu

K – koncentrace

k – konstanta

Aplikace této rovnice na naše texty vypadá takto:

- Psaná čeština

$$\Sigma_P \times \frac{1}{100} K_p = k$$

$$5\,179 \times \frac{1}{100} 15,04 = k$$

$$\mathbf{k = 779}$$

- Mluvená čeština

$$\Sigma_M \times \frac{1}{100} K_m = k$$

$$3\,062 \times \frac{1}{100} 25,2 = k$$

$$\mathbf{k = 788}$$

Pro potvrzení či vyvrácení obecné platnosti naší teorie však bude třeba, aby byla aplikována na další korpusy textů.

7 ROZBORY

7.1 Zájmeno *ten*

Při pohledu na první slova ve frekvenčním slovníku mluvené češtiny nejvíce vyčnívá ukazovací zájmeno *ten*, které se svými 1 368 výskyty téměř dvojnásobně převyšuje druhé slovo *být*, které má 756 výskytů. Rozestup následujících slov již není tak markantní, zpravidla jde o rozdíly v řádu desítek či jednotek. V psaném slovníku tak výrazné postavení žádné heslo nemá a zájmeno *ten* se umístilo v pořadí dle absolutní frekvence na pátém místě s 302 výskyty, tedy o více než tisíc méně. Tento výrazný jev se nyní pokusíme rozebrat a zdůvodnit. Při srovnání s našimi dvěma nejnovějšími frekvenčními slovníky zjistíme, že zájmeno *ten* se ve FSMČ umístil také na prvním místě, a to s takřka stejnou razancí – dvojnásobná četnost oproti druhému slovesu *být*. Při srovnání s FSC zjistíme také téměř stejné výsledky, zájmeno *ten* se vyskytuje na šestém místě, tento fakt napovídá tomu, že se nejedná o nějakou anomálii způsobenou např. špatným výběrem textů, potvrzuje tak relevanci našeho výsledku a celé naší statistiky vůbec.

Pro pochopení a správné zjištění příčiny tak extrémního výskytu zájmena *ten* v mluvených textech je třeba nejdříve vymezit jeho postavení v jazyce jako systému. *Příruční mluvnice češtiny* (PMČ) definuje zájmeno *ten* takto: „Zájmeno *ten* plní funkce ukazovací (v situaci) a odkazovací (v textu). Odkaz je buď anaforický (zpět), nebo kataforický (dopředu). V prvním případě je zájmeno nepřízvučné (*Potkal jsem krásnou dívku a ta dívka...*), ve druhé nepřízvučné (*Vybere si z těch možností, které se mu nabízejí*). Tyto základní funkce mohou být modifikovány dalšími významovými rysy včetně citového zabarvení: *Vežmi si ty nejhorší šaty, co máš; Ta Maru se má; Ty malinké ručičky a ty něžné prstíčky! Toho bláta a té špíny!*“¹¹

V našem souboru mluvených textů se zájmena *ten*, včetně všech jeho modifikací (viz výše), nejvíce užívá ve funkci výplňkového (parazitního) slova, kdy sémantická

¹¹ Grepl, M. a kol.: *Příruční mluvnice češtiny*. Praha 2003, s. 291n.

role je silně potlačena a slovo plní primárně úlohu získání času mluvčího pro konstruování další části výpovědi. V této funkci zájmeno *ten* tak můžeme zařadit k ostatním výplňovým slovům typickým pro mluvený jazyk, o kterých bude pojednáno dále. Neznamená to však, že neplní i jiné funkce, ale zdá se, že tato funkce v našich mluvených textech je dominantní.

Při zkoumání výskytu zájmena *ten* v mluvených textech se přikláníme k tvrzení některých lingvistů, že toto zájmeno před substantivem nabývá podobnou funkci jako určité členy v jiných jazycích. „Již V. Mathesius upozorňuje na to, že oslabení deiktického významu zájmena *ten* postoupilo leckdy hodně daleko, což je patrné z toho, že při překladu do cizí řeči, např. do angličtiny, mnohdy stačí za toto české zájmeno užít určitý člen. Je to zejména u přívlastkového *ten* ve funkci rozlišovací, zvláště u výrazů komparativních, např. *Co dělá ten druhý syn, ten mladší? What does the other son, the younger one, do?*, nebo u superlativů, např. *To je vždycky u mě to první. That is always the first thing with me*, a rovněž před diferencujícími větami relativními, např. *To je to místo, kde jsme byli včera. That is the place we were yesterday.*“¹² V aktuální české jazykové situaci, kdy nejvíce slov přejímáme právě z anglického jazyka a zároveň také s nárůstem aktivních anglicky mluvících Čechů, stále více pronikají do češtiny (zatím však jde o nespisovné tvary užívané zpravidla jen v mluvených projevech) prvky anglické gramatiky. Vzhledem k těmto okolnostem lze předpokládat, že zájmeno *ten* ve funkci členu určitého bude nabývat na síle.

Je však nutné uvést, že hranice mezi jednotlivými funkcemi a významy jsou neostře a vzájemně se prolínají, je tedy téměř nemožné zařazovat jednotlivé konkrétní realizace zájmena *ten* striktně s absolutní platností k jedné funkci, spíše je vhodné rozlišovat, která funkce v konkrétní realizaci převažuje, a je tak dominantní.

¹² Schneiderová, E.: K užívání zájmena *ten* (v přívlastkové pozici) v mluvených projevech. *Naše řeč*, 76, 1993, s. 36–37.

Nápadným výskytem zájmena „ten“ se zabývala Eva Schneiderová v článku *K užívání zájmena ten (v přívlastkové pozici) v mluvených projevech*,¹³ kde rozlišuje čtyři základní významy zájmena *ten*:

„1. Nejčastěji se toto zájmeno využívá jako slovo výplňkové, tedy jako náhražkové slovo, jež má překlenout určitou mezeru vzniklou neznalostí, nejistotou, hledáním vhodného, přesného výrazu. Tuto funkci má zájmeno „ten“ zejména v mluvených projevech dětí, [...].

2. Dále se zájmeno „ten“ objevuje velmi často ve významu připomínacím; připomíná, upozorňuje na to, že uváděná věc je již známá (a to buď objektivně, tedy i posluchači, či jen subjektivně, tedy mluvčímu). Toto často i nadměrné užívání přívlastkového *ten* ve funkci připomínací je opět typické především pro mluvený projev dětí, [...].

3. Přívlastkové zájmeno *ten* ve spojení s pomocným adjektivním určením se také často užívá k rozlišování, plní tedy rozlišovací funkci. Nejvíce příkladů, kde se zájmeno *ten* vyskytuje v této funkci, jsme zaznamenali v rozhovorech při nakupování, [...].

4. Velice zajímavé je užití přívlastkového *ten* rovněž ve funkci expresivní (emocionální), kdy vyjadřuje určité citové zaujetí mluvčího, dále může mít též význam zesilující, zdůrazňovací. Nejvíce příkladů najdeme opět v mluveném projevu dětí či v řeči starších lidí na úřadu či u lékaře, [...].“¹⁴

7.2 Výplňková slova

Obecně známou charakteristikou mluveného jazyka, kterou se výrazně liší od psané podoby, je časté užívání tzv. výplňkových, též parazitních či vycpávkových, slov, jejichž funkce tkví primárně v získání času pro mluvčího při produkci dalších částí výpovědi.

¹³ Schneiderová, E.: K užívání zájmena *ten* (v přívlastkové pozici) v mluvených projevech. *Naše řeč*, 76, 1993, s. 31–37.

¹⁴ Tamtéž, s. 34–36.

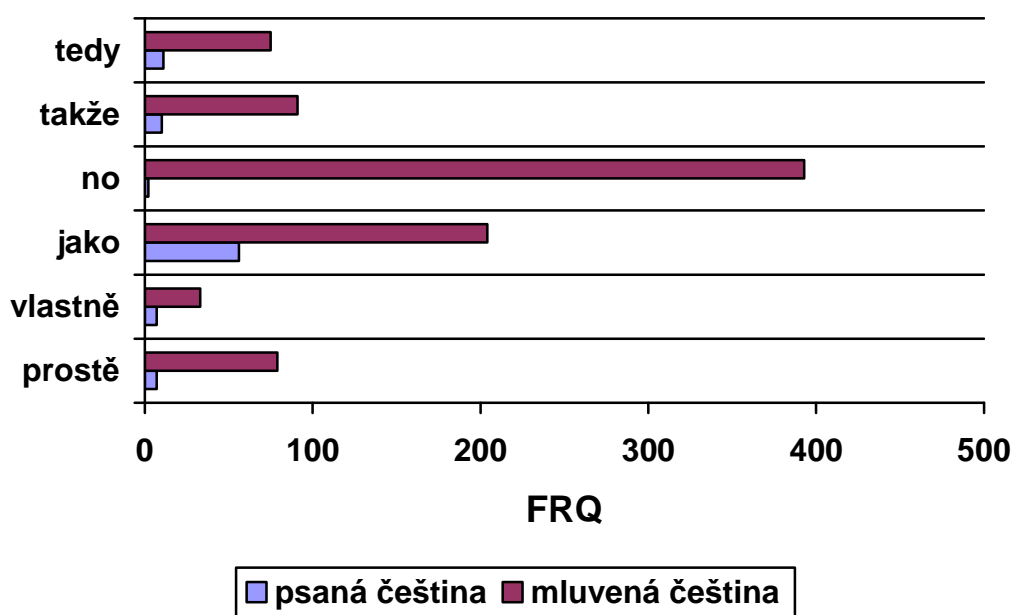
Pro přehlednost uvádíme tabulku s šesti vybranými slovy, která se často užívají jako tzv. výplňková:

slovo	PSANÁ ČEŠTINA			MLUVENÁ ČEŠTINA		
	rank	pořadí	FRQ	rank	pořadí	FRQ
PROSTĚ	72	437.	7	33	35.	79
VLASTNĚ	72	455.	7	61	83.	33
JAKO	28	28.	56	14	14.	204
NO	77	1655.	2	4	4.	393
TAKŽE	69	279.	10	29	29.	91
TEDY	68	236.	11	36	38.	75

Vysvětlivky:

- rank – pořadí dle frekvence, přičemž každá hodnota odpovídá pouze jedné frekvenci, počítá se tedy jen hodnota frekvence, nikoli kvalita jednotky, hodnota ranku je tedy vždy stejná nebo nižší než hodnota pořadí;
- pořadí – hodnota, která určuje pořadí dle absolutní frekvence, přičemž se nepřihlíží k tomu, má-li více jednotek stejnou frekvenci;
- FRQ – hodnota absolutní frekvence, tedy celkový počet výskytů jednotky v korpusu.

Pro lepší znázornění rozdílů výskytů výplňkových slov jsme na základě dat z tabulky vytvořili graf:



Při pohledu na graf je evidentní, že četnost výskytu výplňkových slov je u mluvených textů několikanásobně vyšší než u textů psaných. Příčin tohoto jevu je několik, základním faktorem však je časový limit zásadně ovlivňující charakter mluvených komunikátů. Mluvčí během projevu často naráží na nedostatek času k formulování další výpovědi, užití výplňkových slov je jednou z možností, jak potřebný čas získat. Vedle této možnosti mohou mluvčí pro získání času užít ještě formulační pauzy:

„Vyskytuje se hlavně v odborných projevech a slouží mluvčímu k promýšlení a formulování obsahu sdělení. Pro posluchače představuje nejen oddechový moment, ale je i signálem postupného odvíjení se myšlenek řečníka, že projev vzniká na daném místě, ad hoc. Např. k markantním znakům řečnického stylu D. I. Mendělejeva patřilo to, že své věty prokládal pauzami, když hledal nejpřesnější výraz. Tato pauza se může vyskytovat prakticky na jakémkoliv místě textu, ale neměla by být příliš častá, aby nenarušovala plynulost projevu. Formulační pauzy jsou někdy zaplněny různými hezitačními zvuky. Jsou-li tyto zvuky příliš nápadné a časté, chápe je posluchač jako projev slabší jazykové pohotovosti řečníka nebo jeho nedostatečné připravenosti. Hezitace se někdy považuje za řečnický zlozvyk

(manýru) spočívající v přerušování souvislé řeči ulpíváním na některých samohláskách nebo slabikách.“¹⁵

Výplňková slova však nejsou podpůrným prostředkem pouze na straně mluvčího, ale také na straně posluchače. Tyto vycpávky totiž, stejně jako formulační pauzy, segmentují projev a usnadňují tak recepci textu, užití těchto prostředků také působí v mluveném projevu přirozeně. Edvard Lotko dává výplňková slova do souvislosti s kontaktoými prostředky: „Pro řečníka jsou důležité také lexikální prostředky sloužící k navázání, udržení a posléze k ukončení kontaktu mezi partnery komunikace. Jde o kontaktové prostředky, např. *no, no tak, tedy, nuže, ano, jistě, určitě, víte, všichni víme, podívejte se, poslyšte, děkuji (za pozornost)*... Sem patří i oslovení, které je jedním z prostředků podporujících soustředěnost posluchačů (zahřívá informační kanál). Pokud však řečník používá některé kontaktové prostředky nápadně často a stereotypně, mohou být posluchači vnímány jako řečnický zlovyk (rétorická manýra). Mluvíme zde pak také o slovních vycpávkách, resp. o slovní vatě. Tyto přídatné, parazitující větné prvky jsou typické pro nepřipravené nebo polopřipravené projevy, ve kterých mluvčí nemá jasnou výpovědní perspektivu, srov. ještě *prostě, tak říkajíc, abych tak řekl, prosím pěkně*...“¹⁶

7.3 K (ne)dominanci spojky *a*

Spojka *a* je obecně považována za jednoznačně nejužívanější české slovo, alespoň v psané češtině, neboť již ve FSVS a později i ve FSMČ se tato spojka umístila až na třetím místě. V mluvené češtině bývá první zájmeno *ten* a druhé sloveso *být*, což také platí v případě FSMČ, FSVS i našeho frekvenčního slovníku mluvené češtiny.

V psané češtině se však jak ve FSC, tak ve FSSDTČJ umístila spojka *a* na prvním místě, ale v našem frekvenčním slovníku psané češtiny na druhém. Považujeme tedy

¹⁵ Lotko, E.: *Kapitoly ze současné rétoriky*. Olomouc 2004, s 66.

¹⁶ Tamtéž, s 49.

za vhodné pokusit se objasnit příčiny této odchylky. Jako výchozí podklad použijeme následující tabulku, která přispěje také k přehlednosti dat:

rank	FSC			FSSDTČJ			NÁŠ SLOVNÍK		
	lemma	FRQ	%	lemma	FRQ	%	lemma	FRQ	%
1	A	2792011	100	A	67122	100	V	539	100
2	V/VE	2487420	89	BÝT	43148	64	A	430	78
3	SE	2298438	82	TEN	37280	55	NA	407	76
4	BÝT/BEJT	2042223	73	V/VE	33679	50	BÝT	375	70
5	NA	1643934	58	ON	32496	48	TEN	302	56

Vzhledem k tomu že tyto tři srovnávané frekvenční slovníky byly vypracovány na základě kvantitativně značně rozdílných korpusech, vedle hodnoty absolutní frekvence (FRQ) uvádíme procentní podíl. Tento údaj lépe vyjadřuje kvantitu jednotlivých rozestupů mezi jednotkami. Je vypočítán tak, že hodnota absolutní frekvence prvního lemmatu je 100 %, následující hodnoty pak vyjadřují procentuální podíl z první hodnoty FRQ. Na tomto procentuálním vyjádření můžeme snadno a objektivně srovnávat vzájemné rozestupy jednotlivých jednotek mezi slovníky.

Kromě zmíněných údajů je ještě důležité uvést poměr zastoupení spojky *a* dle funkčních stylů textů v korpusu. Tyto údaje samozřejmě platí jen pro FSC a FSSDTČJ, protože náš slovník psané češtiny byl vypracován výhradně na textech spadajících do publicistického funkčního stylu. Toto srovnání částečně ztěžuje fakt, že zatímco FSC člení texty na beletrii, odbornou literaturu a publicistiku, FSSDTČJ člení texty do osmi skupin (beletrie, básně, literatura pro mládež, dramata, odborná literatura, žurnalistika, vědecká literatura, mluvené projevy). Ke komparaci jsme tak museli těchto osm skupin rozřadit do tří základních a jednotlivé hodnoty sečíst. Značně problematické bylo rozhodnout o přiřazení mluvených projevů (Gottwald, Nejedlý, Burian, Zápotocký), vzhledem k tomu, že tyto projevy vyšly v tisku a vzhledem k jejich autorům jsme považovali za nejvhodnější tyto texty zařadit do publicistiky. Tímto zpracováním dat jsme získali následující tabulku výskytu spojky *a* dle funkčních stylů (uvedené hodnoty jsou v procentech):

	UMĚLECKÝ	ODBORNÝ	PUBLICISTICKÝ
FSC	39	34	27
FSSDTČJ	63	22	15

Z této tabulky je patrné, že spojka *a* se nejméně často vyskytuje v publicistických textech. Jestliže náš frekvenční slovník psané češtiny je založen pouze na publicistických textech, je zřejmé, že právě tento faktor je s největší pravděpodobností primárním činitelem, který odsunul spojku *a* na druhou pozici. Největší zastoupení pak má tato spojka v beletrii, která byla hlavně v případě FSSDTČJ největší částí korpusu. Pokud připustíme, že spojka *a* je na první pozici našich největších frekvenčních slovníků zejména proto, že má dominantní postavení v beletristických textech, pak obecné povědomí o tom, že spojka *a* je jednoznačně nejčastějším českým slovem, musíme označit za nesprávné. Beletrii totiž jen těžko můžeme považovat za reprezentativní texty psané češtiny, neboť umělecké texty jsou příliš specifické a jsou vždy osobitým vyjádřením autora. Texty uměleckého funkčního stylu jsou nepřírozené, tvořené velmi úzkou a vyhraněnou skupinou lidí. Beletrie nemá žádná pravidla ani normu, dokonce se zde ani nemůžeme mluvit o úzu, užívá se v ní široká škála výrazových prostředků, často se objevují také fiktivní mluvené repliky, tyto texty jsou natolik rozmanité, že jen těžko v nich můžeme najít stabilní ukazatele, proto také existují frekvenční slovníky jednotlivých autorů. V případě publicistiky tomu tak není, texty vykazují mnohem větší konzistenci, pokud budeme srovnávat články různých autorů nebo různých periodik, výsledky lexikální analýzy budou vykazovat minimální odchylky. Ještě větší míru konzistence vykazují odborné texty, které jsou v absolutní opozici vůči beletrii, to znamená maximální dodržování normy a úzu, minimální odlišnosti textů v závislosti na autorovi. Jedinou spojitost s beletrií můžeme spatřovat v tom, že okruh autorů je také značně úzký.

Troufáme si tak tvrdit, že spojka *a* rozhodně nemá tak dominantní postavení v psané češtině, jak by se mohlo zdát. Toto tvrzení kromě již zmíněného opíráme o data z první tabulky, kde právě u FSSDTČJ druhá jednotka (sloveso *být*) dosahuje pouhých 64 % hodnoty jednotky první, u FSC již rozdíl tak markantní není, druhá jednotka (předložka *v/ve*) dosahuje 89 % hodnoty frekvence spojky *a*. Jestliže

v našem slovníku je právě předložka *v/ve* na prvním místě před spojkou *a*, je velmi pravděpodobné, že při omezení zastoupení textů uměleckého funkčního stylu ve FSČ by se předložka *v/ve* mohla také teoreticky umístit na první pozici.

7.4 Komparace hodnot s FSVS

Zajímavé je srovnání prvních deseti slov mluvené češtiny dle FSVS a našeho slovníku. Přestože FSVS byl vypracován na základě korpusu, kde mluvenou češtinu představovaly jen odborné texty (tedy zpravidla předem připravené – tedy na hranici psanosti-mluvenosti), v komparaci s výsledky psané češtiny byly zjištěny hodnoty indikující elementární rozdíly psané a mluvené češtiny.

rank	FSVS	NÁŠ SLOVNÍK
1	TEN	TEN
2	BÝT	BÝT
3	A	A
4	ŽE	NO
5	V/VE	TAK
6	TEDY	ŽE
7	TAK	JÁ
8	KTERÝ	MÍT
9	MÍT	V
10	NA	ON

Přes výše zmíněný charakter mluvených textů v korpusu FSVS můžeme ve srovnávací tabulce vidět nápadnou shodu prvních slov s naším vypracovaným slovníkem. Pořadí prvních tří slov je totožné, tři další stejná slova se pak vyskytují v obou slovnících na zbylých různých pozicích.

Z tabulky vyplývá, že námi získané hodnoty nejsou nahodilé, toto srovnání tak podporuje relevanci našich údajů, a to i přesto, že nedošlo k úplné shodě prvních

deseti pozic dle absolutní frekvence. Úplná shoda by byla krajně nepravděpodobným jevem, a to hned z několika důvodů, např.:

- značně rozdílný charakter excerpovaných textů co do funkčních stylů v korpusech;
- různý kvantitativní rozsah korpusů (FSVS – 100 000 slov, náš slovník – 20 000 slov);
- různé zásady zpracování lingvistických dat, především v chápání lexikální jednotky jsou pravděpodobné určité rozdíly;
- různé stáří dat (FSVS byl vydán již v roce 1983).

Nyní se pokusíme vysvětlit příčiny alespoň některých rozdílů v tabulce. Velmi nápadný je výskyt slova *tedy/teda* na šesté pozici dle FSVS, v našem slovníku obsadilo až třicátou šestou pozici. Naopak v našem slovníku hned čtvrtou pozici obsadilo slovo *no*, zatímco ve FSVS se vůbec v prvních deseti slovech neumístilo. Máme za to, že ač je vysvětlení složité, neboť dochází k interferenci několika faktorů, primární roli zde hraje posunutá sémantika obou slov a zní vyplývající i jejich funkce v textu.

Obě slova zpravidla v mluvených textech ztrácejí svůj primární význam a sémanticky se vyprazdňují, zastávají tak většinou funkci tzv. výplňkových či vycpávkových slov (viz výše). Z tohoto důvodu se obě tyto jinak různé jednotky stávají v podstatě jednotkou totožnou. Vzhledem k již zmíněnému rozdílnému charakteru jednotlivých korpusů tak tyto odchylky ve frekvenci chápeme jako různé vyjádření téhož, přičemž volba různých podob je podmíněna především odlišností prostředí, v nichž byly výpovědi realizovány, tudíž na základě objektivních slohotvorných činitelů. Pokud je naše hypotéza platná, znamená to, že při prvotním pohledu na tabulku zřetelně rozdílné hodnoty jsou po detailnějším rozboru mnohem menší, možná i nulové. Tento fakt by tak jen potvrzoval správné zpracování našeho slovníku a ukazoval na obecná specifika mluvené češtiny, v tomto případě problematiku tzv. vycpávkových slov (viz výše).

Dalším vážným rozdílem se může zdát předložka *na*, která obsadila v FSVS desátou pozici, zatímco v našem slovníku se nevyskytuje vůbec. Zde je řešení velmi

jednoduché, neboť předložka *na* obsadila hned první neuvedenou pozici, tedy jedenáctou, rozdíl je tak zcela zanedbatelný.

Za zmínku zde ještě stojí objasnění zájmena *já*, které obsadilo v našem slovníku sedmou pozici, zatímco ve FSVS se vůbec mezi prvními deseti slovy neumístilo. Tento jev je způsoben jednak tím, že v našem korpusu byly zpravidla dialogické texty narozdíl od FSVS, kde se jednalo především o texty monologické, jednak tím, že v našem případě šlo o texty spontánní, kdežto u FSVS se jednalo o texty předem připravené.

7.5 Fatické prostředky v mluvených textech

Důležitým charakteristickým rysem mluvených textů, zejména dialogických, je enormní výskyt kontaktních prostředků, jejichž funkce tkví ve vytváření a udržování kontaktu mezi jednotlivými účastníky komunikace. Je zřejmé, že tyto prostředky se v psaných monologických komunikátech vyskytují jen velmi omezeně či vůbec. Zaměřili jsme se tak na četnost výskytu několika fatických prostředků v našem mluveném korpusu, výsledky jsou uvedeny v následující tabulce:

kontaktní prostředek	rank	FRQ
NO	4	393
JO	13	226
ANO	67	26
AHA	74	19

Jednotlivé výrazy jsme do tabulky zařadili na základě textu Olgy Müllerové, která uvádí částice *ano*, *jo*, *no*, *aha* jako nejčastější kontaktní prostředky v dialogických komunikátech. Kromě uvedených částic sem Müllerová řadí také zvuk *hm*, který jsme do analýzy nezařadili, neboť se jedná o různě realizovaný zvuk, který tak není lexikální jednotkou. Přesto je třeba jeho častý výskyt v souvislosti s fatickými prostředky zmínit. K částici *no* je třeba uvést, že jsme ji již výše zařadili mezi

tzv. výplňková slova, neboť tento prostředek plní více funkcí, které se vzájemně překrývají, to také vysvětluje jeho velmi vysokou hodnotu absolutní frekvence.

Z tabulky vyplývá, jak časté jsou jednotlivé prostředky v korpusu užity. Je třeba zmínit, že zde jsme nepřistoupili ke komparaci s hodnotami psaného korpusu, neboť tyto prostředky se v psaných a zpravidla monologických textech nevyskytují, v našem psaném korpusu se *ano* vyskytuje jen čtyřikrát a *no* pouze dvakrát, *aha* a *jo* se pak neobjevuje vůbec.

7.6 Komparace našich slovníků s FSC, FSSDTČJ a FSMČ

Poslední část sedmé kapitoly věnujeme srovnání našich dvou slovníků s třemi nejvýznamnějšími frekvenčními slovníky – FSC, FSSDTČJ, FSMČ. Považujeme totiž za vhodné vzhledem k předchozím rozborům a celkovému charakteru této práce přehledně vedle postavit alespoň část dat těchto prací. Dosáhneme tak nejen konfrontace našich slovníků s ostatními výše uvedenými, ale také mezi nimi vzájemně, navíc budou postaveny vedle sebe jak práce zpracovávající psanou češtinu, tak i mluvenou.

Na základě těchto požadavků jsme vytvořili tabulku, která obsahuje prvních 30 jednotek každého slovníku. Kvůli maximální snaze o přímé srovnání jsme nemohli do tabulky zařadit hodnoty absolutní frekvence ani žádné jiné. Přesto se domníváme, že tato tabulka bude zajímavým doplněním této práce, neboť je možné v ní sledovat mnohé charakteristiky, zejména ve vztahu psané a mluvené češtiny. Také, jak již bylo zmíněno, by měla být tato tabulka oporou předchozích rozborů.

POŘADÍ	PSANÁ ČEŠTINA			MLUVENÁ ČEŠTINA	
	FSC	NÁŠ SLOVNÍK	FSSDTČJ	FSMČ	NÁŠ SLOVNÍK
1.	A	V	A	TEN	TEN
2.	V/VE	A	BÝT	BÝT	BÝT
3.	SE	NA	TEN	A	A
4.	BÝT/BEJT	BÝT	V/VE	ŽE	NO
5.	NA	TEN	ON	JÁ	TAK
6.	TEN	ŽE	NA	MÍT	ŽE
7.	S/SE	Z	ŽE	NO	JÁ
8.	ŽE	KTERÝ	S/SE	V	MÍT
9.	Z/ZE	ON	Z/ZE	NA	V
10.	KTERÝ/KTEREJ	I	KTERÝ	JAKO	ON
11.	O/VO	DO	MÍT	TAK	NA
12.	MÍT	S	JÁ	ALE	TY
13.	DO	ZA	K/KE/KU	TAM	JO
14.	ON/VON	O	DO	TEDA	JAKO
15.	I	MÍT	I	PROSTĚ	ALE
16.	K/KE/KU	ROK	ALE	MYSLET	TAM
17.	ROK/LÉTA	ALE	SVŮJ	NEBO	NE
18.	PRO	K	JAKO	TŘEBA	S
19.	SVŮJ	ČLOVĚK	O	S	DO
20.	ALE	UŽ	TAK	PROTOŽE	JAK
21.	ZA	PRO	CO	UŽ	UŽ
22.	TENTO	SVŮJ	ZA	ON	CO
23.	MOCT/MOCI	PO	SE	JO	VĚDĚT
24.	JAKO	SE	TENTO	MOCT	KDYŽ
25.	JEHO	MOCT	MOCI	CO	NEBO
26.	PO	OD	JEN	ONI	TAKOVÝ
27.	JÁ	PODLE	PO	TAKOVÝ	TADY
28.	TAK	JAKO	ABY	Z	ŘÍKAT
29.	ONI	JEN	JAK	DO	TAKŽE
30.	OD/ODE	TAK	JENŽ	DÍTĚ	Z

8 ZIPFOVY ZÁKONY

George Kingsley Zipf (1902–1950) byl americký profesor lingvistiky na Harvardské univerzitě, snažil se jazykovědu více přiblížit exaktním přírodním vědám, své vědecké bádání soustředil na aplikaci matematiky při zkoumání systému přirozeného jazyka. Zipf tak byl jedním z průkopníků kvantitativní lingvistiky, a přestože se často setkával s nepochopením, dnes je uznávanou osobností nejen v samotné jazykovědě, ale také v dalších vědách, zejména pak statistice nebo ekonomii.

V této kapitole budeme aplikovat tzv. Zipfovy zákony na našich dvou slovnících, pokusíme se tak podpořit či vyvrátit jejich platnost v českém jazyce a se zřetelem k základním cílům práce budeme míru adaptability Zipfových zákonů srovnávat z hlediska psané a mluvené češtiny.

Nejznámější Zipfův zákon charakterizuje vztahy mezi frekvencí a pořadím slova, kromě tohoto tzv. prvního Zipfova zákona se pak pokusíme aplikovat ještě další dva: tzv. druhý a třetí Zipfův zákon.

8.1 První Zipfův zákon

Tento jednoznačně nejznámější a nejuniverzálnější zákon aplikovatelný snad ve všech vědách postihuje vztah mezi frekvencí a pořadím nejružnějších entit – v rámci lingvistiky slov. Zipf při statistických analýzách různých textů došel k poznání, že součin frekvence slova a jeho ranku je konstantní. Základní rovnice je tak následující:

$$f \times r = k$$

f – frekvence (FRQ)

r – rank

k – konstanta

Z této základní rovnice tak vyplývá, že slovo s největší četností se vyskytuje v textu dvakrát častěji než druhé nejčastější slovo, třikrát častěji než třetí, čtyřikrát častěji než čtvrté. Tato řada pokračuje stejným způsobem do nekonečna. Důležité pro pochopení podstaty tohoto vztahu je uvědomění možnosti jeho aplikace i v jiných nelingvistických statistikách. Jako názornou ukázkou zde uvedeme některé poznatky z analýz německého statistika Waltera Krämera, který zjistil, že: „Nejrozšířenější řečí na zemi (čínštinou) hovoří dvojnásobně tolik lidí než druhou nejčastější (hindštinou), třikrát tolik jako třetí nejrozšířenější (angličtinou) a čtyřikrát tolik jako čtvrtou nejrozšířenější (španělštinou). Německý týdeník s největším nákladem (Bild am Sonntag) má dvojnásobný náklad než druhý nejsilnější týdeník (Das Beste) a třikrát větší než třetí nejsilnější (Stern) a čtyřikrát větší než čtvrtý nejsilnější (Focus).“¹⁷ Krämer Zipfův zákon se stejnými výsledky uplatnil na vztahu největších německých měst a počtem jejich obyvatel.

S vysvětlením platnosti Zipfova zákona přišel držitel Nobelovy ceny za ekonomii Herbert Simon, který dokázal, že „[...] platí takovéto rozdělení pro přirozeně rostoucí systémy, u nichž jednak růst nezávisí na velikosti jednotky a jednak vychází sám ze sebe. A tak také největší město USA (New York) je dvojnásobně tak velké než druhé největší (Chicago), třikrát tak velké jako třetí největší (Los Angeles), čtyřikrát tak velké jako čtvrté největší, atd. Ale největší město Anglie nebo Rakouska není vůbec dvakrát tak velké jako druhé největší nebo třikrát tak velké jako třetí největší. Protože dnešní Anglie nebo Rakousko jsou zbytky někdejších koloniálních nebo mnohonárodnostních mocností a nikoliv organicky rostoucími útvary. Pokud by se však k Rakousku připojilo ještě Maďarsko, Česko a Slovensko, nebo k Anglii Indie a Kanada, pak by se Zipfův zákon opět uplatnil: Vídeň je dvakrát tak velká jako Budapešť a třikrát tak velká jako Praha, a podobný vztah bychom našli a bez problémů dokázali i mezi Londýnem, Kalkatou, Torontem a Bombají a dalšími městy států někdejšího britského impéria.“¹⁸

Zipfův zákon tak s největší pravděpodobností můžeme spolehlivě aplikovat na všechny přirozené systémy lidské společnosti, přirozený jazyk nevyjímaje. Zipfův zákon tak exaktně potvrzuje základní teze, a to, že přirozený jazyk je společenský jev

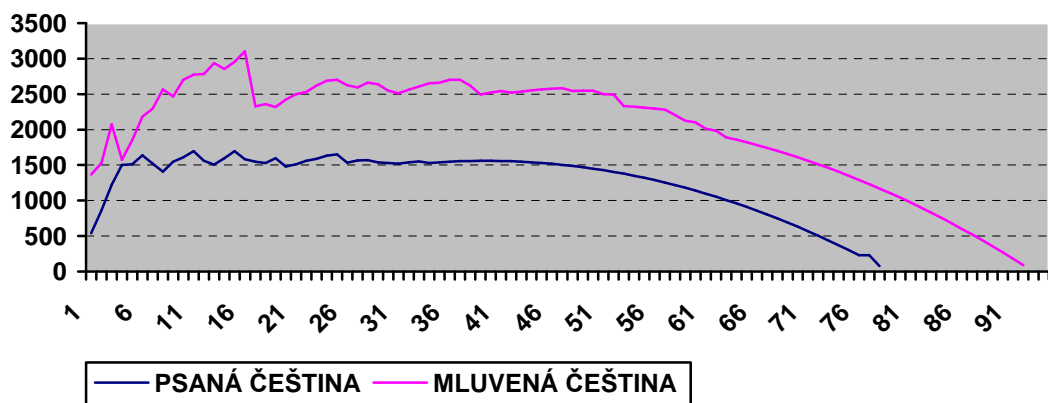
¹⁷ Krämer, W.: *Statistika do vesty*. Praha 2005, s. 119–120.

¹⁸ Tamtéž, s. 121–122.

a systém znaků. Otázkou je, zda Zipfův zákon půjde aplikovat na češtinu, u které je vzhledem k výrazně flektivnímu charakteru relativně obtížné stanovit základní jednotku (viz výše), také se nabízí otázka, do jaké míry má rušivý vliv ovlivňování češtiny jinými jazyky (zvláště němčiny). Nyní se pokusíme aplikovat Zipfovu rovnici na naše slovníky psané a mluvené češtiny:

PSANÁ ČEŠTINA											
r	f	k	r	f	k	r	f	k	r	f	k
1	539	539	21	72	1512	41	38	1558	61	18	1098
2	430	860	22	71	1562	42	37	1554	62	17	1054
3	407	1221	23	69	1587	43	36	1548	63	16	1008
4	375	1500	24	68	1632	44	35	1540	64	15	960
5	302	1510	25	66	1650	45	34	1530	65	14	910
6	228	1638	26	59	1534	46	33	1518	66	13	858
7	217	1519	27	58	1566	47	32	1504	67	12	804
8	176	1408	28	56	1568	48	31	1488	68	11	748
9	172	1548	29	53	1537	49	30	1470	69	10	690
10	161	1610	30	51	1530	50	29	1450	70	9	630
11	154	1694	31	49	1519	51	28	1428	71	8	568
12	130	1560	32	48	1536	52	27	1404	72	7	504
13	116	1508	33	47	1551	53	26	1378	73	6	438
14	114	1596	34	45	1530	54	25	1350	74	5	370
15	113	1695	35	44	1540	55	24	1320	75	4	300
16	99	1584	36	43	1548	56	23	1288	76	3	228
17	91	1547	37	42	1554	57	22	1254	77	2	231
18	85	1530	38	41	1558	58	21	1218	78	1	78
19	84	1596	39	40	1560	59	20	1180			
20	74	1480	40	39	1560	60	19	1140			

MLUVENÁ ČEŠTINA											
r	f	k	r	f	k	r	f	k	r	f	k
1	1368	1368	24	112	2688	47	55	2585	70	23	1610
2	765	1530	25	108	2700	48	53	2544	71	22	1562
3	691	2073	26	101	2626	49	52	2548	72	21	1512
4	393	1572	27	96	2592	50	51	2550	73	20	1460
5	370	1850	28	95	2660	51	49	2499	74	19	1406
6	364	2184	29	91	2639	52	48	2496	75	18	1350
7	328	2296	30	85	2550	53	44	2332	76	17	1292
8	321	2568	31	81	2511	54	43	2322	77	16	1232
9	274	2466	32	80	2560	55	42	2310	78	15	1170
10	270	2700	33	79	2607	56	41	2296	79	14	1106
11	252	2772	34	78	2652	57	40	2280	80	13	1040
12	232	2784	35	76	2660	58	38	2204	81	12	972
13	226	2938	36	75	2700	59	36	2124	82	11	902
14	204	2856	37	73	2701	60	35	2100	83	10	830
15	197	2955	38	69	2622	61	33	2013	84	9	756
16	194	3104	39	64	2496	62	32	1984	85	8	680
17	137	2329	40	63	2520	63	30	1890	86	7	602
18	131	2358	41	62	2542	64	29	1856	87	6	522
19	122	2318	42	60	2520	65	28	1820	88	5	440
20	121	2420	43	59	2537	66	27	1782	89	4	356
21	119	2499	44	58	2552	67	26	1742	90	3	270
22	115	2530	45	57	2565	68	25	1700	91	2	182
23	114	2622	46	56	2576	69	24	1656	92	1	92



Při pohledu do tabulek a na graf je patrné, že první Zipfův zákon pro češtinu neplatí obecně. Podle zjištěných dat můžeme říci, že pravidlo konstantního součinu frekvence a ranku slova platí pro češtinu jen ve středním frekvenčním pásmu. V pásmu slov s nejvyšší frekvencí je zřejmý progresivní vzestup hodnoty, v oblasti nízké frekvence pak můžeme sledovat pozvolný pravidelný sestup hodnoty, pouze ve středním pásmu křivka osciluje v rozmezí, které můžeme označit za konstantní, neboť amplitudy v tomto pásmu dosahují od nulové (průměrné) přímky maximální odchylky kolem 100. Na základě hodnot získaných aplikací prvního Zipfova zákona můžeme potvrdit jeho platnost pro češtinu, ovšem s omezením pouze na střední frekvenční pásmo.

Pokud jde o diferenci výsledných hodnot v psané a mluvené češtině, musíme konstatovat, že opozice psanost-mluvenost zde nehraje žádnou roli. Tento fakt je nejlépe patrný z vypracovaného grafu, kde průběh křivek má relativně stejný charakter.

8.2 Druhý Zipfův zákon

Tzv. druhý Zipfův zákon vyjadřuje vztah mezi frekvencí a počtem různých slov, která jí odpovídají, touto rovnicí:

$$a \times b^2 = k$$

a – suma počtu slov s určitou frekvencí

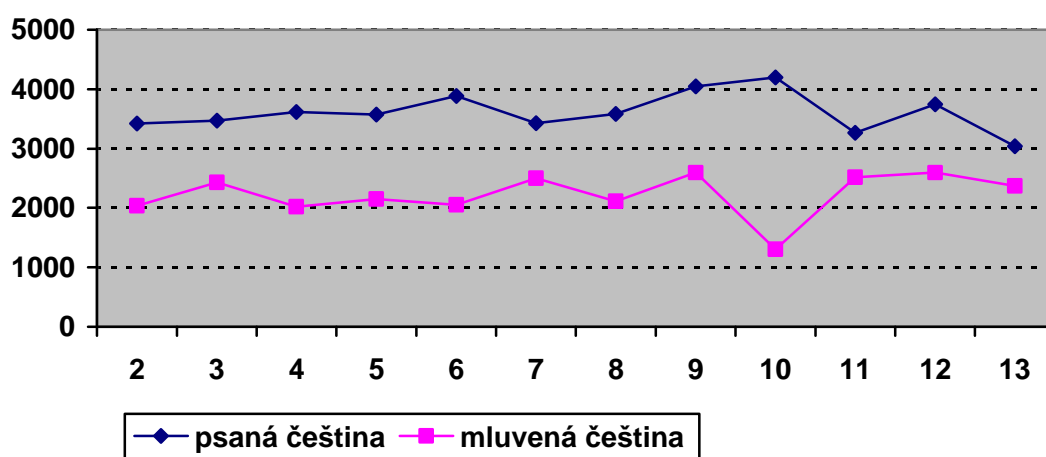
b – frekvence

k – konstanta

Zipf tak říká, že součin počtu slov s určitou frekvencí a hodnotou frekvence umocněnou na druhou je konstantní. Nutno uvést, že platnost této rovnice samotný Zipf omezuje pouze na slova s nízkou frekvencí, která však představují značnou část

slovníku. Při aplikaci rovnice na naše slovníky jsme zařadili jednotky s frekvencí 2–12. Vypočtené hodnoty uvádíme v tabulce a následně v grafu:

PSANÁ ČEŠTINA			MLUVENÁ ČEŠTINA		
a	b	k (a×b ²)	a	b	k (a×b ²)
856	2	3424	508	2	2032
386	3	3474	270	3	2430
226	4	3616	126	4	2016
143	5	3575	86	5	2150
108	6	3888	57	6	2052
70	7	3430	51	7	2499
56	8	3584	33	8	2112
50	9	4050	32	9	2592
42	10	4200	13	10	1300
27	11	3267	21	11	2541
26	12	3744	18	12	2592
18	13	3042	14	13	2366



Na základě zjištěných hodnot uvedených v tabulce a zpracovaných v grafu můžeme potvrdit platnost druhého Zipfova zákona, neboť skutečně můžeme sledovat konstantní charakter hodnot. Jak již bylo však výše uvedeno, samotný Zipf platnost této zákonitosti vztahoval pouze na oblast slov s nízkou frekvencí.

Z hlediska opozice psanosti-mluvenosti pak můžeme konstatovat stejný závěr jako v případě prvního zákona, a to irelevanci této opozice ve vztahu ke sledované charakteristice.

8.3 Třetí Zipfův zákon

Tzv. třetí Zipfův zákon postihuje vztah frekvence slova a počtů jeho významů. Zipf tvrdí, že nejvíce užívaná slova jsou více polysémní než slova užívaná zřídka. Základní rovnice definující tento vztah je:

$$m / \sqrt{f} = k$$

m – počet významů slova

f – frekvence slova

k – konstanta

Při aplikaci této rovnice na naše dva frekvenční slovníky jsme zjistili, že třetí Zipfův zákon není platný pro češtinu, a to jak psanou, tak mluvenou. Máme za to, že třetí Zipfův zákon spíše jen poukazuje na obecně známý fakt, že jednotka s více významy (funkcemi) je užívaná více než kdyby měla významů méně. Tento fakt však svým charakterem spadá spíše do oboru statistiky než lingvistiky. Přesto, že je tento fakt pravdivý, v žádném případě to neznamená, že by mezi nejfrekventovanějšími jednotkami byly pouze či většinou slova s více sémantickými významy. To stejné pak naopak platí u jednotek s nízkou frekvencí, kde se vyskytují i slova s více významy. Jak již bylo zmíněno, považujeme třetí Zipfův zákon pouze za snahu nějak definovat jednoduché logické pravidlo, které však není relevantní pro jazyk jako ucelený systém.

9 ZÁVĚR

Všechny cíle stanovené v úvodu této práce byly splněny. Sestavili jsme psaný a mluvený korpus, na jejichž základě jsme vytvořili frekvenční slovníky a provedli lexikální analýzu. Výstupní data těchto výpočtů byly podrobeny vzájemným komparacím, jejichž výsledky jsme pak analyzovali v následujících kapitolách.

Nejprve jsme soustředili zájem na tři elementární parametry – rozsah, rozptýlení a koncentraci lexika. Z hlediska opozice psanost-mluvenost jsme největší odchylky zaznamenali v rozsahu a koncentraci. V případě rozsahu lexika jsme jako hlavní příčiny těchto diferencí stanovili zejména tyto faktory: připravenost-nepřipravenost, charakter podavatele, téma textů, oficiálnost-soukromost, explicitnost, míru abstrakce a organizovanost.

Při analýze výsledků koncentrace jsme zjistili, že tento parametr je značně determinován rozsahem lexika a roli tak zde hrají stejné faktory. Dále jsme zjistili, že mezi celkovým počtem jednotek ve slovníku a součtem slov deseti nejfrekventovanějších jednotek je vztah nepřímé úměry – čím větší je počet jednotek ve slovníku, tím menší je počet deseti nejfrekventovanějších slov. Na základě tohoto zjištění jsme také stanovili rovnici $\Sigma_1 \times N_1^{10} = \Sigma_2 \times N_2^{10}$, která tento vztah vyjadřuje.

Naopak jako takřka inertní vzhledem k opozici psanost-mluvenost se jeví rozptýlení lexika. Výsledné hodnoty nevykazovaly takřka žádné odchylky. Zdá se tedy, že rozptýlení lexika s největší pravděpodobností je spíše charakteristikou jazyka jako konkrétního přirozeného systému znaků sloužícího ke komunikaci jednoho jazykového společenství, a ne jako parametr s korelací s psanou, či mluvenou formou jazyka.

Po analýzách uvedených základních parametrů jsme na základě komparace psaného a mluveného frekvenčního slovníku stanovili několik jevů majících značně odlišnou četnost výskytů (frekvenci). Tyto jevy jsme podrobili rozborům a snažili se příčiny jejich existence interpretovat.

Nejprve jsme analyzovali značně odlišnou hodnotu frekvence zájmena *ten*, které má velmi vysoký výskyt v mluvené češtině. Přestože faktorů určujícím tento fakt je více, jako základní činitel jsme určili jeho primární funkci v mluvených

komunikátech. Nejvíce se jej užívá ve funkci výplňkového (parazitního) slova, kdy sémantická role je silně potlačena a slovo plní primárně úlohu získání času mluvčího pro konstruování další části výpovědi. Druhý nejvýznamnější činitel pak spatřujeme v tom, že toto zájmeno před substantivem nabývá podobnou funkci jako určité členy v jiných jazycích.

Dále jsme pozornost zaměřili na tzv. výplňková slova, která jsou typickým rysem mluvených textů, jejichž funkce tkví primárně v získání času pro mluvčího při produkci dalších částí výpovědi. Tento jev jsme demonstrovali na grafickém srovnání frekvence šesti nejčastěji užívaných výplňkových slov (prostě, jako, no, takže, vlastně, tedy).

Při komparaci našeho frekvenčního slovníku s FSC a FSSDTČJ jsme zjistili, že v našem slovníku má spojka *a* narozdíl od ostatních až druhou pozici v absolutní frekvenci. Příčiny této odchylky jsme našli především v odlišné struktuře analyzovaných korpusů, kdy jak u FSC, tak u FSSDTČJ značnou část korpusů představovaly beletristické texty, které považujeme z několika důvodů za nereprezentativní vzorek psané češtiny. Právě u těchto textů byla zjištěna velmi vysoká frekvence spojky *a*, která se tak dostala na první pozici obou slovníků. Dominantní roli spojky *a* tak považujeme minimálně za diskutabilní.

Náš frekvenční slovník mluvené češtiny jsme pak srovnávali s FSVS, kdy jsme vedle sebe postavili deset nejfrekventovanějších jednotek. V této komparaci jsme našli zásadní odchylky v užívání částic *tedy* a *no*, kde *tedy* mělo vysokou frekvenci v našem slovníku a *no* zase ve FSVS. Při rozboru jsme konstatovali, že pravděpodobná příčina těchto odchylek je způsobena pohnutou sémantikou obou slov, kdy tato různá dvě slova vyjadřovala v podstatě totéž v kooperaci s funkcí komunikátu, v němž se tyto jednotky realizují.

Jednou z největších diferencí mezi psanou a mluvenou češtinou je enormní užívání fatických prostředků v mluvených textech. Z hodnot odečtených z našich slovníků jsme tento fakt potvrdili. Je zřejmé, že tyto prostředky, jejichž funkce tkví ve vytváření a udržování kontaktu mezi jednotlivými účastníky komunikace, se v mluvených dialogických textech objevují velmi často. Naopak v psaných monologických komunikátech se vyskytují jen velmi omezeně nebo vůbec.

V závěrečné části práce jsme aplikovali na naše frekvenční slovníky tzv. Zipfovy zákony. Jednak jsme chtěli ověřit míru jejich platnosti v současné češtině, jednak jsme chtěli zjistit, zda existuje nějaká korelace mezi jejich platností a psanou a mluvenou formou jazyka. Zatímco u prvního a druhého Zipfova zákona jsme určitou platnost pro češtinu našli, třetí zákon jsme pak označili pro češtinu jako irelevantní.

10 ANOTACE

Příjmení a jméno autora: Kubát Miroslav

Název katedry a fakulty: Katedra bohemistiky Filozofické fakulty Univerzity Palackého v Olomouci

Název diplomové práce: Rozsah, rozptýlení a koncentrace lexika v psané a mluvené češtině

Vedoucí diplomové práce: PhDr. Ondřej Bláha, Ph.D.

Počet znaků: 73 566

Počet příloh: 2

Počet titulů použité literatury: 21 (+ 4 internetové odkazy)

Klíčová slova: mluvená čeština, psaná čeština, kvantitativní lingvistika, lexikální analýza, frekvence slov, Zipfovy zákony, statistika

Keywords: spoken Czech, written Czech, quantitative linguistics, lexical analysis, word frequency, Zipf's laws, statistics

Ключевые слова: устный чешский язык, письменный чешский язык, квантитативная лингвистика, лексический анализ, частота слова, законы Зипфа, статистика

Resumé

Tato práce na základě lexikální analýzy psaných a mluvených textů zjišťuje elementární distinkce současné psané a mluvené češtiny. Sledovány jsou zejména tři základní charakteristiky lexikální statistiky – rozsah, rozptýlení a koncentrace lexika. Získaná data jsou podrobena rozborům a konfrontována s podobnými pracemi. Část práce se věnuje také aplikaci tzv. Zipfových zákonů a ověřování jejich platnosti nejen v současné češtině jako celku, ale zejména ve vztahu k psanosti a mluvenosti. Součástí práce je také frekvenční slovník psané a mluvené češtiny.

Summary

This work on the basis of lexical analysis of written and spoken texts detects elemental distinctions contemporary written and spoken Czech. Mainly three basic characteristics of lexical statistics are monitored – extent, dispersion and concentration of vocabulary. Obtained data are analyzed and confronted with similar works. Part of the work deals with the application of so called Zipf's laws and the verification of their validity, not only in the current Czech as a whole, but particularly in relation to written and spoken form. Part of this work is the frequency dictionary of written and spoken Czech.

Резюме

Данная работа на основе лексического анализа письменных и устных текстов обнаруживает основные различительные признаки современного письменного и устного чешского языка. Наблюдаем в основном три основные характеристики лексической статистики – диапазон, дисперсия и концентрация лексики. Полученные данные анализируются и сопоставляются с аналогичными работами. Часть работы посвящена применению так называемых законов Зипфа и их верификации не только в современном чешском языке в целом, но особенно в отношении письменной и устной форм. Частью данной работы является частотный словарь письменного и устного чешского языка.

11 ZDROJE

LITERATURA

- BOGOCZOVÁ, Irena a kol.: *Tváře češtiny*. Ostrava, Ostravská univerzita 2000, 266 s.
- ČERMÁK, František a kol.: *Frekvenční slovník češtiny*. Praha, Nakladatelství Lidové noviny 2004, 596 s.
- ČERMÁK, František a kol.: *Frekvenční slovník mluvené češtiny*. Praha, Karolinum 2007, 512 s.
- ČERMÁK, František a kol.: *Slovník Karla Čapka*. Praha, Nakladatelství Lidové noviny 2007, 715 s.
- ČMEJRKOVÁ, Světla: Mluvenost a psanost. In: *Přednášky z doktorandských dnů*. Praha, Univerzita Karlova 2006, s. 55–79.
- DAVIDOVÁ, Dana (ed.): *K diferenciaci současného mluveného jazyka*. Ostrava, Ostravská univerzita 1995. 219 s.
- DEJMEK, Bohumír: Frekvence slov a tvarů substantiv a sloves v běžně mluveném jazyce. *Slovo a slovesnost*, 63, 2002, s. 15–25.
- GREPL, Miroslav a kol.: *Příruční mluvnice češtiny*. Praha, Nakladatelství Lidové noviny 2003, 800 s.
- JELÍNEK, Jaroslav – BEČKA, Josef V. – TĚŠITELOVÁ, Marie: *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha, SPN 1961, 588 s.
- JÍLEK, Viktor: *Žurnalistické texty jako výsledek působení jazykových a mimojazykových vlivů*. Olomouc, Univerzita Palackého 2009, 122 s.
- KARLÍK, Petr – NEKULA, Marek – PLESKALOVÁ, Jana (eds.). *Encyklopedický slovník češtiny*. Praha, Nakladatelství Lidové noviny 2002. 604 s.
- KRÄMER, Walter: *Statistika do vesty*. Praha, Baronet 2005, 124 s.
- LOTKO, Edvard: *Kapitoly ze současné rétoriky*. Olomouc, Univerzita Palackého 2004, 202 s.

- MÜLLEROVÁ, Olga a kol.: *Mluvená čeština v autentických textech*. Jinočany, H + H 1992, 236 s.
- PETR, Jan a kol.: *Mluvnice češtiny (2)*. Praha, Academia 1986, 536 s.
- SCHNEIDEROVÁ, Eva: K užívání zájmena ten (v přívlastkové pozici) v mluvených projevech. *Naše řeč*, 76, 1993, s. 31–37.
- TĚŠITELOVÁ, Marie a kol.: *O češtině v číslech*. Praha, Academia 1987, 208 s.
- TĚŠITELOVÁ, Marie a kol.: *Psaná a mluvená odborná čeština z kvantitativního hlediska*. Praha, Ústav pro jazyk český ČSAV 1983, 147 s.
- TĚŠITELOVÁ, Marie: *Kvantitativní lingvistika*. Praha, SPN 1987, 187 s.
- TĚŠITELOVÁ, Marie: *Otázky lexikální statistiky*. Praha, Academia 1974, 292 s.
- VAŠÁK, Pavel: *Metody určování autorství*. Praha, Academia 1980, 236 s.

INTERNETOVÉ ODKAZY

- <http://ucnk.ff.cuni.cz>
- <http://www.ujc.cas.cz>
- <http://www.mytranslate.com/wordfrequency.htm>
- <http://www.mfdnes.cz>

12 PŘÍLOHA

FREKVENČNÍ SLOVNÍK PSANÉ ČEŠTINY

Pozn.: Uvedeny jsou pouze jednotky s frekvencí 5 a vyšší. V prvním sloupci je uvedena hodnota ranku (r), v druhém frekvence (f).

r	f	lemma
1	539	V
2	430	A
3	407	NA
4	375	BÝT
5	302	TEN
6	228	ŽE
7	217	Z
8	176	KTERÝ
9	172	ON
10	161	I
11	154	DO
12	130	S
13	116	ZA
14	114	O
15	113	MÍT
16	99	ROK
17	91	ALE
18	85	K
19	84	ČLOVĚK
20	74	UŽ
21	72	PRO
22	71	SVŮJ
23	69	PO
24	68	SE
25	66	MOCT
26	59	OD
27	58	PODLE
28	56	JAKO
	56	JEN
	56	TAK
29	53	CO
30	51	CENA
31	49	NOVÝ
32	48	JEDEN
33	47	AŽ
	47	DVA
	47	JAK
	47	KDYŽ
	47	PRAHA
34	45	VŠAK
35	44	CHTÍT
	44	ONI

	44	STÁT (substantivum)
	44	VŠECHEN
36	43	JEŠTĚ
	43	NEBO
37	42	ČESKÝ
	42	DALŠÍ
38	41	JÁ
	41	MY
39	40	ÚTOK
40	39	VELKÝ
41	38	NEŽ
42	37	VČERA
43	36	EVROPA
	36	PŘED
44	35	MUSET
	35	PŘI
45	34	PROCENTO
46	33	DÁT
	33	ŘÍCT
47	32	PROTI
48	31	DOSTAT
	31	PAK
49	30	AL
	30	KORUNA
50	29	ANI
	29	PŘÍŠTÍ
	29	ZEMĚ
51	28	DNES
	28	U
52	27	SÁM
53	26	VLAK
54	25	ABY
	25	ZAČÍT
55	24	DUBAI
	24	MEZI
	24	PROTO
	24	TAKÉ
	24	TŘI
56	23	JÍT
	23	SLUŽBA
	23	SVOBODNÝ
57	22	KDO
	22	KDY
	22	NĚKTERÝ

	22	PRVNÍ
	22	ŘÍKAT
	22	SOUD
	22	TAKOVÝ
	22	VEĎĚT
58	21	BEZ
	21	CELÝ
	21	INFORMACE
	21	JEHO
	21	JEJICH
	21	KVŮLI
	21	MU
	21	PŘIJÍT
	21	TOTIŽ
59	20	BANKA
	20	KAŽDÝ
	20	KDE
	20	MLUVČÍ
	20	OLOMOUC
	20	PLYN
	20	TAJNÝ
	20	VÍCE
60	19	AUTOR
	19	DAŇ
	19	DÍTĚ
	19	DOBA
	19	IRÁCKÝ
	19	JINÝ
	19	MÍT
	19	NAPŘÍKLAD
	19	NAVÍC
	19	ODS
	19	PĚT
	19	PROTOŽE
	19	ZPRÁVA
61	18	ANÍ
	18	ČESKO
	18	ČI
	18	DOBŘÝ
	18	MOŽNÝ
	18	TAM
	18	VLÁDA
62	17	ČÁST
	17	DLUH
	17	CHŘÍPKA
	17	MĚSTO
	17	MILION
	17	MUŽ
	17	OASIS
	17	PLÁN
	17	TEĎ
	17	TŘEBA
	17	TÝDEN
	17	UVÉST
	17	VĚTŠINA

	17	VYSOKÝ
	17	VYSVĚTLIT
63	16	AMERICKÝ
	16	AUTO
	16	ČEKAT
	16	ČTYŘI
	16	DEMJANJUK
	16	IRÁK
	16	KRIZE
	16	NÁŠ
	16	PŘÍPAD
	16	STÁTNÍ
	16	TERORISTA
	16	VY
	16	ZATÍM
	16	ŽÁDNÝ
64	15	DEN
	15	DESET
	15	DÍVKÁ
	15	PRÁVĚ
65	14	BUDOVA
	14	DÁLE
	14	E
	14	KNIHA
	14	KRAJ
	14	PŘES
	14	SOUDCE
	14	STÁT (verbum)
	14	SVĚT
	14	TŘI
66	13	ASI
	13	DŮM
	13	EMIRÁT
	13	FIRMA
	13	CHÁLID
	13	KDYBY
	13	KOMUNISTA
	13	MALÝ
	13	MÍSTO
	13	MNOHO
	13	NĚKOLIK
	13	NIKDY
	13	PACIENT
	13	POČÍTAT
	13	POKUD
	13	POLITIK
	13	PŘEDEVŠÍM
	13	TÉMĚŘ
67	12	AMERIČAN
	12	CESTA
	12	COŽ
	12	DOKONCE
	12	DOPRAVA
	12	JESTLI
	12	JISTÝ

	12	LÉKAŘ
	12	MINISTR
	12	NEW
	12	OLOMOUCKÝ
	12	PÁR
	12	POZDĚ
	12	PRÁCE
	12	PRAŽSKÝ
	12	PRÝ
	12	PŘIZNAT
	12	RÁDIO
	12	SESTRA
	12	STAROSTA
	12	STARÝ
	12	STRANA
	12	VIDĚT
	12	YORK
	12	ZŘEJMĚ
	12	ŽIVOT
68	11	AKCE
	11	DOKÁZAT
	11	DOST
	11	DRUHÝ
	11	HODNOTA
	11	JEDINÝ
	11	MĚSÍC
	11	MILIARDA
	11	NE
	11	NĚJAKÝ
	11	NYNÍ
	11	PENÍZE
	11	POD
	11	POMOC
	11	PROBLÉM
	11	SEDM
	11	SMRT
	11	SVĚTOVÝ
	11	TEDY
	11	TEHDY
	11	TREST
	11	USA
	11	VÁLKA
	11	VČETNĚ
	11	VŮZ
	11	ZAPLATIT
	11	ZÁŘÍ
69	10	AUTOMOBILKA
	10	ČASTO
	10	DĚLAT
	10	EVROPSKÝ
	10	HODINA
	10	CHYSTAT
	10	KAPR
	10	KOLEM
	10	KOLIK

	10	KONEC
	10	LÉK
	10	LZE
	10	MF
	10	MOC
	10	MYSLET
	10	NĚKDO
	10	NEMOCNICE
	10	NIC
	10	NIKDO
	10	NÍZKÝ
	10	OBA
	10	OSOBNÍ
	10	OTÁZKA
	10	PLATIT
	10	POSLANEC
	10	POSLEDNÍ
	10	PREZIDENT
	10	PROCES
	10	PŘITOM
	10	PSÁT
	10	ROZPOČET
	10	SADDÁM
	10	STEJNÝ
	10	ŠÉF
	10	TAKŽE
	10	TISÍC
	10	TRH
	10	UKÁZAT
	10	VIZE
	10	VŮBEC
	10	ZÁKON
	10	ZÍSKAT
	10	ZNAMENAT
70	9	BAGDÁD
	9	BĚHEM
	9	BIS
	9	CENTRUM
	9	DAVID
	9	DIVADLO
	9	DOSUD
	9	DRÁHA
	9	DŘÍV
	9	DVA TISÍCE JEDNA
	9	FAKULTA
	9	FINANČNÍ
	9	HODNĚ
	9	INFUZE
	9	JAKÝ
	9	JEDENÁCT
	9	JMÉNO
	9	KAJDA
	9	KILOMETR
	9	KOSTELEC
	9	KSM

	9	LETENKA
	9	LETOS
	9	LEVNÝ
	9	MAIL
	9	MINISTERSTVO
	9	NABÍDNOUT
	9	NAJÍT
	9	NĚCO
	9	NĚMECKÝ
	9	NEMOVITOST
	9	OBĚD
	9	PLÁNOVAT
	9	PÓDIUM
	9	POLICIE
	9	POPRVÉ
	9	PRAVDA
	9	PROČ
	9	PROJEKT
	9	PŘÍLIŠ
	9	REGIONÁLNÍ
	9	SKUTEČNOST
	9	SPOLEČNOST
	9	STEJNĚ
	9	TRAŤ
	9	TROCHU
	9	TVRDIT
	9	VYPADAT
	9	ZASE
	9	ŽIDOVSKÝ
71	8	ABŮ
	8	AŤ
	8	ATENTÁT
	8	BAROKNÍ
	8	ČAS
	8	ČÍNA
	8	ČÍNSKÝ
	8	ČSSD
	8	DESÍTKA
	8	DLOUHÝ
	8	DOZORCE
	8	HLAVNĚ
	8	CHVÍLE
	8	JARO
	8	JENŽE
	8	JEZDIT
	8	KOEFICIENT
	8	KOUPIT
	8	LETADLO
	8	METR
	8	MÍSTNÍ
	8	MLUVIT
	8	MODERNÍ
	8	MORAVA
	8	MOŘE
	8	MŮJ

	8	MÝTO
	8	NABÍDKA
	8	NÁMĚSTÍ
	8	NÁVRH
	8	OBJEVIT
	8	OSM
	8	PATNÁCT
	8	PODOBNÝ
	8	POLOVINA
	8	POSTAVIT
	8	POTŘEBOVAT
	8	PRÁVO
	8	PROSINEC
	8	PŘÍMO
	8	REGION
	8	ŘÍJEN
	8	SICE
	8	SKUPINA
	8	SNÍŽENÍ
	8	SPÍŠ
	8	STÁLE
	8	SYSTÉM
	8	USÁMA
	8	VINĚTA
	8	VODA
	8	ZÁROVEŇ
	8	ZNAČKA
	8	ZNALOST
	8	ZNOVU
	8	ZŮSTAT
72	7	ALESPOŇ
	7	AMBASÁDA
	7	ATTA
	7	BIN
	7	BUĎ
	7	BYT
	7	BÝVALÝ
	7	BYZNYS
	7	ČECH
	7	ČESKO
	7	ČÍSLO
	7	DOJÍT
	7	DOLAR
	7	DUBAJSKÝ
	7	DVA TISÍCE DVA
	7	HÁČEK
	7	HLAVNÍ
	7	HROZIT
	7	JAN
	7	JASNÝ
	7	JENŽ
	7	KAMERA
	7	KAPELA
	7	KLAUS
	7	KLESAT

	7	LÁDIN
	7	LITERÁRNÍ
	7	MINULÝ
	7	NAD
	7	NAKONEC
	7	NAOPAK
	7	NÁZOR
	7	ODPOVĚDĚT
	7	PADNOUT
	7	PATŘÍT
	7	PEDAGOGICKÝ
	7	PETR
	7	POČET
	7	POHLED
	7	POLICISTA
	7	POROTCE
	7	PRASEČÍ
	7	PRAVÝ
	7	PROSTĚ
	7	PRŮBĚH
	7	PŘIPRAVENÝ
	7	ROZJET
	7	RUKA
	7	SAMOZŘEJMĚ
	7	SCHŮZKA
	7	SILNÝ
	7	SOBIBOR
	7	STŘEDNÍ
	7	STUDIE
	7	ŠEST
	7	ŠPATNÝ
	7	TENTO
	7	TERORISTICKÝ
	7	URČITĚ
	7	VEČER
	7	VINIT
	7	VLASTNĚ
	7	VOJENSKÝ
	7	VÝSLEDEK
	7	VŽDYCKY
	7	ZAČÁTEK
	7	ZASTUPITEL
	7	ZDA
	7	ZMĚNA
	7	ZRUŠIT
73	6	ARABSKÝ
	6	BRITSKÝ
	6	CÍTIT
	6	CLICK4SKY
	6	ČLOVĚK
	6	DÁLNIČNÍ
	6	DÁVKA
	6	DEVĚT
	6	DHABI
	6	DODAVATEL

	6	DOSTÁVAT
	6	DŮVOD
	6	DVA TISÍCE DVANÁCT
	6	DVA TISÍCE SEDM
	6	EPIDEMIE
	6	FUNGOVAT
	6	GUANTÁNAMO
	6	HOSPODÁŘSKÝ
	6	INTERNET
	6	INTERNETOVÝ
	6	JANA
	6	JEDNAT
	6	JINAK
	6	JISTĚ
	6	KABINA
	6	KAM
	6	KOCOUREK
	6	KROMĚ
	6	KVĚTEN
	6	LISTOPAD
	6	MAJETEK
	6	MÉDIUM
	6	MINARET
	6	MUHAMMAD
	6	NÁKLAD
	6	NECHAT
	6	NEMOC
	6	NEWYORSKÝ
	6	OBAVA
	6	OBDOBÍ
	6	ODSTARTOVAT
	6	PODARIT
	6	PODÍVAT
	6	PODMÍNKA
	6	PODOBŇĚ
	6	POMÁHAT
	6	POPLATEK
	6	POŘÁD
	6	POST
	6	POTĚ
	6	POTVRDIT
	6	PRODAT
	6	PROFIL
	6	PROPADAT
	6	PRŮMYSL
	6	PRŮMYSLOVÝ
	6	PŘECE
	6	PŘESTO
	6	PŘIDAT
	6	PŘINĚST
	6	PŮSOBIT
	6	RAKETA
	6	REAGOVAT
	6	REPUBLIKA
	6	ROPA

	6	ROVNOU
	6	ROZHOVOR
	6	RŮST
	6	ŘÍZENÍ
	6	SAAB
	6	SALONARIUM
	6	SKONČIT
	6	SLOVO
	6	SNAD
	6	SNAŽIT
	6	SOUDNÍ
	6	SOUPRAVA
	6	STAVĚT
	6	STRACH
	6	ŠAJCH
	6	ŠKOLA
	6	ŠPION
	6	TADY
	6	TĚŽKÝ
	6	THE
	6	TRÁM
	6	TRANSPORTÉR
	6	TŘETÍ
	6	TURISTA
	6	TVÁŘ
	6	TVOŘIT
	6	ÚŘAD
	6	VÁCLAV
	6	VEŘEJNOST
	6	VĚZENÍ
	6	VLÁDNÍ
	6	VOLBY
	6	VRÁTIT
	6	VYBRAT
	6	VYJÍT
	6	VÝŠKA
	6	VZRŮST
	6	ZAVEDENÍ
	6	ZBRAŇ
	6	ZDROJ
	6	ZJISTIT
	6	ZNÁMÝ
	6	ZNÁT
74	5	ANALYTICI
	5	ARMÁDA
	5	BACKSTREET
	5	BÁT
	5	BEZPEČNOSTNÍ
	5	BLÍZKÝ
	5	BRZY
	5	BURZA
	5	CEMPÍREK
	5	ČÍŇAN
	5	ČLEN
	5	DĚTSKÝ

	5	DEVADESÁT
	5	DLOUHO
	5	DOBŘE
	5	DOHODNOUT
	5	DRAHÝ
	5	DRASLÍK
	5	DŮKAZ
	5	DVA TISÍCE TŘI
	5	DVACET
	5	DVĚMA
	5	ELEKTRONICKÝ
	5	FANOUŠEK
	5	FORD
	5	HLAS
	5	HROZBA
	5	HUDEBNÍ
	5	HUSAJN
	5	CHLORID
	5	INVESTIČNÍ
	5	INVESTOR
	5	ISLÁMSKÝ
	5	JAKL
	5	JANČURA
	5	JOHN
	5	KLÍČOVÝ
	5	KLUB
	5	KOČÁRNÍK
	5	KULISA
	5	LONI
	5	MÉNĚ
	5	MĚSÍČNĚ
	5	MÍRNĚ
	5	MNICHOV
	5	MODRÝ
	5	MOTOR
	5	MYSTIFIKACE
	5	NADÁLE
	5	NEDÁVNO
	5	NOC
	5	NOVELA
	5	NULA
	5	OBAMA
	5	OBLAST
	5	OBŽALOBA
	5	OČKOVÁNÍ
	5	ODVOLAT
	5	OKNO
	5	OPATŘENÍ
	5	OPERA
	5	OSLOVIT
	5	OSOBA
	5	OSUD
	5	OTEC
	5	OZNÁMIT
	5	PÁD

	5	POBYT
	5	POSLAT
	5	POUZE
	5	POUŽÍVAT
	5	PRACOVAT
	5	PROCHÁZKA
	5	PŘIJET
	5	PŘIPRAVIT
	5	PŘÍSTAV
	5	PŮL
	5	PŮVODNĚ
	5	PŮVODNÍ
	5	RÁD
	5	RATH
	5	ROČNĚ
	5	ROZTOK
	5	RUSKÝ
	5	ŘADA
	5	ŘIDIČ
	5	SEJÍT
	5	SÍN
	5	SITUACE
	5	SKUTEČNĚ
	5	SNÍŽIT
	5	SOCIÁLNÍ
	5	SPOJENÍ
	5	SPOJENÝ
	5	SRDCE
	5	STAVEBNÍ
	5	STO
	5	ŠEST
	5	ŠEVČÍK
	5	ŠUBERT
	5	TAKZVANÝ
	5	TAMNÍ
	5	TEHDEJŠÍ
	5	TĚLO
	5	TERMÍN
	5	TERORISMUS
	5	TRPĚT
	5	UDÁLOST
	5	UKRAJINEC
	5	UNIVERZITA
	5	ÚSPORA
	5	VÁNOČNÍ
	5	VÁŠ
	5	VĚC
	5	VĚK
	5	VENEK
	5	VĚRIT
	5	VĚST
	5	VYJÁDRIT
	5	VÝROBCE
	5	VYŠETŘOVATEL
	5	VZHLED

	5	VŽDYŽ
	5	ZABÍT
	5	ZAČÍNAT
	5	ZÁJEM
	5	ZÁKAZ
	5	ZAKÁZKA
	5	ZÁPADNÍ
	5	ZDRAVÍ
	5	ZDRAVOTNÍ
	5	ZDRAVÝ
	5	ZLOČINEC
	5	ZMĚNIT
	5	ZPĚT
	5	ZRANIT
	5	ZROVNA
	5	ZVEŘEJNIT
	5	ZVÝŠENÍ
	5	ZVÝŠIT
	5	ŽENA
	5	ŽID

FREKVENČNÍ SLOVNÍK MLUVENÉ ČEŠTINY

Pozn.: Uvedeny jsou pouze jednotky s frekvencí 5 a vyšší. V prvním sloupci je uvedena hodnota ranku (r), v druhém frekvence (f).

r	f	lemma
1	1368	TEN
2	765	BÝT
3	691	A
4	393	NO
5	370	TAK
6	364	ŽE
7	328	JÁ
8	321	MÍT
9	274	V
10	270	ON
11	252	NA
12	232	TY
13	226	JO
14	204	JAKO
15	197	ALE
16	194	TAM
17	137	NE
18	131	S
19	122	DO
20	121	JAK
21	119	UŽ
22	115	CO
23	114	VĚDĚT
24	112	KDYŽ
25	108	NEBO
26	101	TAKOVÝ
27	96	TADY
28	95	ŘÍKAT
29	91	TAKŽE
	91	Z
30	85	ONI
31	81	VY
32	80	MOCT
	80	O
33	79	PROSTĚ
34	78	SE
35	76	ZA
36	75	TEDY
	75	MUSET
37	73	JEŠTĚ
	73	MY
38	69	TENTO

39	64	PROTOŽE
40	63	I
41	62	TAKY
42	60	ABY
	60	TŘEBA
43	59	NĚCO
	59	NĚJAKÝ
	59	PRO
	59	ŘÍCT
44	58	CHTÍT
45	57	TEĎ
46	56	ASI
	56	JENOM
47	55	JÍT
48	53	K
	53	KTERÝ
49	52	JEDEN
	52	MI
50	51	ZASE
51	49	DOBŘE
	49	PAK
52	48	NIC
53	44	DÁT
	44	POTOM
54	43	DOBŘÝ
	43	DĚLAT
	43	STO
55	42	MYSLET
56	41	DVA
	41	PRÁVĚ
57	40	HODNĚ
	40	JESTLI
	40	TŘI
	40	ŠKOLA
58	38	U
59	36	VŠECHNO
60	35	JET
	35	VŠECHEN
	35	ČLOVĚK
61	33	PO
	33	VLASTNĚ
62	32	CELÝ
63	30	DEN
	30	MOC

	30	NÁŠ
	30	UDĚLAT
64	29	DÍTĚ
	29	NĚJAK
	29	VELKÝ
	29	VLASTNOST
65	28	BÝT
	28	DRUHÝ
66	27	ÚPLNĚ
67	26	ANO
	26	CHODIT
	26	DALŠÍ
	26	VŮL
	26	VĚC
	26	ŠPUNDOVA
68	25	AŽ
	25	TAKHLE
69	24	KDYBY
	24	OD
70	23	KDE
	23	PRVNÍ
	23	VIDĚT
	23	VŮBEC
71	22	JINÝ
	22	PŘIJÍT
	22	TEĎKA
72	21	LÁSKA
	21	ŽÁDNÝ
73	20	ANI
	20	DOSTAT
	20	JEHO
	20	KAŽDÝ
	20	OPRAVDU
	20	STÁT (verbum)
74	19	AHA
	19	FAKT
	19	PŘI
	19	SAMOZŘEJMĚ
	19	UMĚT
	19	VZÍT
75	18	DNESKA
	18	JAKÝ
	18	KOUPIT
	18	PRÁCE
76	17	MĚSTO
	17	NĚKDE
	17	POKUD
	17	TENHLE
	17	TISÍC
	17	VODA
77	16	HEZKÝ
	16	MALÝ
	16	OBEC
	16	ROK
	16	RUKA

	16	SEM
	16	SPÍŠ
	16	SVŮJ
78	15	BRÁT
	15	DOMA
	15	MEZI
	15	POTŘEBOVAT
	15	TĚŽKÝ
	15	VÁŠ
79	14	CÍTIT
	14	DALEKO
	14	HOLKA
	14	KDY
	14	MÁLO
	14	NĚKDO
	14	STROJ
	14	ÚŘAD
80	13	BOLET
	13	FIRMA
	13	JINAK
	13	MANŽELSTVÍ
	13	NEŽ
	13	NORMÁLNĚ
	13	PÁD
	13	PŘIJET
	13	PŘÍKLAD
	13	STRAŠNĚ
	13	URČITÝ
	13	VÝSTAVBA
	13	VŽDYCKY
	13	ZÁKON
81	12	ASPOŇ
	12	KDO
	12	MÍSTO
	12	NĚKDY
	12	POŽÁR
	12	PĚKNÝ
	12	PŘED
	12	RÁD
	12	RÁNO
	12	SLOVESO
	12	STRANA
	12	VZEJÍT
	12	VŠAK
	12	ZEPTAT
	12	ZNÁT
	12	ZÁKLADNÍ
	12	ČÁST
	12	ŽIVOT
82	11	DOST
	11	DRAHÝ
	11	HODINA
	11	KOLEM
	11	KVĚTEN
	11	LÍBIT

	11	MOŽNÝ
	11	ODBORNÝ
	11	PODÍVAT
	11	POŘÁD
	11	PRAVDA
	11	PROBLÉM
	11	PROSIT
	11	PĚT
	11	SEDMÝ
	11	SÁM
	11	TÝDEN
	11	UVAŽOVAT
	11	ZATÍM
	11	ZARÍZENÍ
	11	ZÁSTUPCE
83	10	BEZ
	10	DOBA
	10	JEDNOU
	10	KNÍŽKA
	10	KONEC
	10	NAPSAT
	10	NOHA
	10	POKOJ
	10	RÁDIO
	10	TŘETÍ
	10	UČIT
	10	VYPADAT
	10	ZAČÍT
	10	ČAS
84	9	BAVIT
	9	DEVADESÁT
	9	DIKTAFON
	9	DLOUHO
	9	JENŽE
	9	JEŽÍŠ
	9	LET
	9	MOŽNÁ
	9	MŮJ
	9	NĚKTERÝ
	9	OSM
	9	PADESÁT
	9	PANÍ
	9	PENÍZE
	9	POŘÁDEK
	9	PROTI
	9	PROČ
	9	PRÝ
	9	PŘES
	9	RŮZNÝ
	9	STEJNĚ
	9	STŘEDNÍ
	9	UČITELKA
	9	VRÁTIT
	9	VÝROBA
	9	ZASTUPITELSTVO

	9	ČESKÝ
	9	ČEŠTINA
	9	ČTYŘI
	9	ČÍST
	9	ŠEDESÁT
	9	ŠEST
85	8	CELKEM
	8	DOKTOR
	8	DŮM
	8	FINANČNÍ
	8	HROZNÝ
	8	JAZYK
	8	JEJICH
	8	JEN
	8	KOLIK
	8	KŮŇ
	8	MILION
	8	NECHAT
	8	NIKDO
	8	NOVÝ
	8	OKRESNÍ
	8	PAN
	8	PLUS
	8	POD
	8	PODLE
	8	POČKAT
	8	PROCENTO
	8	PROTO
	8	PÁTÝ
	8	PŘESNĚ
	8	SEDMNÁCT
	8	SMĚT
	8	SNAD
	8	TAKÉ
	8	TEPLO
	8	VYJADŘOVAT
	8	ZNAMENAT
	8	ČEKAT
	8	ŠPATNĚ
86	7	ADAPTÉR
	7	AKORÁT
	7	ANEBO
	7	BRUNTÁL
	7	BYDLET
	7	CENA
	7	DEVATENÁCT
	7	DEVÁTÝ
	7	DNES
	7	DNEŠNÍ
	7	DRŽET
	7	DVACET
	7	DVAKRÁT
	7	DÁVAT
	7	FURT
	7	HNED

	7	HÁDAT
	7	JIŽ
	7	JMÉNO
	7	KABELOVÝ
	7	KOMISE
	7	LEŽET
	7	LÍNÝ
	7	MANŽEL
	7	NAPADNOUT
	7	OPRAVA
	7	PODOBNĚ
	7	PORADIT
	7	POUŽÍVAT
	7	PRACOVAT
	7	PROSTŘEDEK
	7	PTÁT
	7	PŘÍSUDEK
	7	RADA
	7	SLOVO
	7	STÁTNICE
	7	TAHLE
	7	TROŠKU
	7	UMŘÍT
	7	URČITĚ
	7	VSTUP
	7	WALKMAN
	7	ZAPLATIT
	7	ZKOUŠKA
	7	ZÁJEM
	7	ZŮSTAT
	7	ÚTERÝ
	7	ČERVEN
	7	ŘEŠIT
	7	ŠKOLKA
87	6	AŤ
	6	BABIČKA
	6	BÍLÝ
	6	CÍL
	6	DESET
	6	DEVĚT
	6	DOCELA
	6	DOPADNOUT
	6	DOTACE
	6	DVACÁTÝ
	6	DĚJEPIS
	6	DRÍV
	6	HODNOCENÍ
	6	HYGIENA
	6	INVESTIČNÍ
	6	JASNÝ
	6	JEDNÁNÍ
	6	KANÁL
	6	KLUK
	6	KOLEJ
	6	KONČIT

	6	KVŮLI
	6	NAJÍT
	6	NEMOCNÝ
	6	NYNÍ
	6	NÁZOR
	6	PENZION
	6	PODNIK
	6	PSÁT
	6	PŘÍDAVNÝ
	6	PŘÍPADNĚ
	6	PŮL
	6	RODINA
	6	SEDĚT
	6	SICE
	6	SPOLEČNOST
	6	SYSTÉM
	6	TATÍNEK
	6	TENKRÁT
	6	TRPNÝ
	6	TÁTA
	6	VELMI
	6	VLAK
	6	VRT
	6	VSTÁVAT
	6	VÁLKA
	6	VĚŘIT
	6	ZDE
	6	ZJISTIT
	6	ZNÁMÝ
	6	ZÁDA
	6	ZŘIZOVATEL
	6	ÚNOR
	6	ČEZ
	6	ČTRNÁCT
	6	ŘADA
88	5	ALTERNATIVA
	5	AP
	5	AUTOR
	5	BRZY
	5	BUDOVA
	5	BÝVALÝ
	5	BŘEZEN
	5	CHRÁNĚNÝ
	5	DOHROMADY
	5	DOJÍT
	5	DOKONCE
	5	DUBEN
	5	DVEŘE
	5	DÍVAT
	5	DĚKOVAT
	5	EKVALIZÉR
	5	HRÁT
	5	INVESTOVAT
	5	INŽENÝR
	5	JEDNAT

	5	JEZDIT
	5	KONCEPCE
	5	KORUNA
	5	KOČÁREK
	5	KRK
	5	KROK
	5	LIPICÁN
	5	LOKALITA
	5	MATERIÁL
	5	MEZHOLEZ
	5	MINUTA
	5	MĚSTSKÝ
	5	MĚSÍC
	5	NIKDY
	5	PAPÍR
	5	PATNÁCT
	5	PODMÍNKA
	5	POSLEDNÍ
	5	POVÍDAT
	5	POČÍTAT
	5	PROSINEC
	5	PROSTOR
	5	PŘESTO
	5	PŘÍPAD
	5	PŘÍTEL
	5	PŘÍŠTÍ
	5	REVIZOR
	5	ROD
	5	RODIČ
	5	ROZUMĚT
	5	SOUHLASIT
	5	STANOVISKO
	5	STAV
	5	TAŤKA
	5	TEHDY
	5	TKALCOVNA
	5	TYP
	5	TÁŇA
	5	TÝDNĚ
	5	UŽÍVAT
	5	VESNICE
	5	VEČER
	5	VHODNÝ
	5	VOLNÝ
	5	VŮZ
	5	VYSOKÝ
	5	VÉZT
	5	VĚTŠINOU
	5	ZABALOVAČKA
	5	ZASTAVIT
	5	ZAČNOUT
	5	ZDRHAT
	5	ZKUSIT
	5	ZPŮSOB

	5	ZÁMĚR
	5	ČASTO
	5	ČI
	5	ČTVRTÝ
	5	ČTYŘICET
	5	ČÍSLO
	5	ŠATY
	5	ŠKOLSKÝ
	5	ŠPATNÝ
	5	ŽIVÝ
	5	ŽÁDOST