Univerzita Karlova v Praze

Filozofická fakulta

Ústav Českého národního korpusu

Studijní program: filologie

Studijní obor: matematická lingvistika

Michal Křen

Diachronní srovnání synchronních korpusů
Diachronic comparison of synchronic corpora

Disertační práce

Školitel: prof. PhDr. František Čermák, DrSc.

2012

Prohlášení

Prohlašuji, že jsem disertační práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 29. března 2012

Abstrakt

Práce představuje metodu pro diachronní srovnání synchronních korpusů zachycujících blízké stavy jazyka. Cílem práce je především zhodnotit možnosti a meze detekce vývojových tendencí v jazyce na materiálu synchronních psaných korpusů řady SYN. Metodologicky jde o corpus-driven přístup založený na statistickém vyhodnocení rozdílů mezi normalizovanými průměrnými redukovanými frekvencemi lemmat a lexikálních kombinací. Metoda je aplikována v několika variantách na různě definované subkorpusy korpusu SYN a podrobně vyhodnocena.

Provedené srovnání ztěžuje především vliv složení jednotlivých korpusů a provázanost změn v jazyce se změnami společenskými. Protože neumíme spolehlivě odlišit zárodky diachronních posunů od přirozeně existující synchronní variability, je statisticky zjištěná významnost frekvenčních rozdílů jednotlivých výrazů zpětně ověřována na korpusech a interpretace výsledků korigována znalostí jejich přesného složení.

Závěry jsou založeny především na publicistice, která je z psaného jazyka nejvíce otevřená změnám. Změny v jazyce publicistiky lze charakterizovat jako tematický odklon od původní politické a ekonomické orientace směrem k tématům týkajícím se praktického života a využívání volného času spojený se zvyšující se neformálností, která způsobuje posuny ve frekvencích některých slovních druhů, frekvenční nárůst řady lemmat z jádra slovní zásoby, vzrůstající podíl významově oslabených sloves, obměnu některých šablonovitých spojení atd.

K přínosům práce patří také vyhodnocení složení korpusů řady SYN, zvláště reprezentativních korpusů SYN2000, SYN2005 a SYN2010. Výsledkem jsou praktická doporučení ke změnám v konceptu reprezentativnosti, kategorizaci textů a složení korpusových dat, která tvoří cennou zpětnou vazbu pro budování dalších korpusů této řady.

Klíčová slova

synchronní korpusy, diachronní srovnání, lexikální frekvence, jazykový vývoj, jazyková variabilita, složení korpusu, reprezentativnost

Abstract

The thesis presents a method for diachronic comparison of synchronic corpora that reflect language of very close time periods. Its primary aim is the assessment of possibilities and limitations of language change detection based on the synchronic written SYN-series corpora. The approach is corpus-driven, based on a statistical evaluation of differences among normalized average reduced frequencies of lemmata and lexical combinations. There are several variants of the method applied on various subcorpora of corpus SYN and their results examined in detail.

Difficulty of the comparison lies in the influence of corpus composition and the interconnection of changes in language with changes in society. As it is not easy to distinguish the signs of diachronic shift from naturally existing synchronic variability, the statistically discovered significance of frequency differences is additionally verified by querying the base corpora. The interpretation of the results is also adjusted by the knowledge of their exact composition.

The conclusions are based mainly on the newspapers as a written text type that is most receptive to the changes. The changes can be characterized as a thematic diversion from the original political and economical orientation of the newspapers towards real-life and free-time topics associated with increasing informality of the language. The informality has an impact on shifts in part-of-speech frequencies, frequency increase of a number of core vocabulary lemmata, growing share of semantically weak verbs, substitution of some conventional expressions etc.

The thesis also contributes to the evaluation of composition of the SYN-series corpora, especially the representative corpora SYN2000, SYN2005 and SYN2010. As a result, a number of practical improvements of the concept of representativeness, text categorization and data composition are formulated. The suggestions constitute a valuable feedback for compilation of future SYN-series corpora.

Keywords

synchronic corpora, diachronic comparison, lexical frequencies, language change, language variability, corpus composition, representativeness

Obsah

1 Úvod

Kvantitativní studium jazyka je v současné době nemyslitelné bez použití jazykových korpusů jako základního zdroje informací. Výstavbou jazykových korpusů a kontinuálním mapováním češtiny se dlouhodobě a cíleně zabývá především Ústav Českého národního korpusu. Od jeho založení vzniklo mnoho jazykových korpusů různého druhu a zaměření, vyvážených žánrově, sociolingvisticky nebo nijak nevyvážených, psaných i mluvených synchronních, a také korpus diachronní. Při jejich vzniku však přirozeně vyvstávaly nejenom otázky zabývající se vztahem mezi korpusem a jazykovou realitou, růzností jednotlivých žánrů a jejich ideálním zastoupením v budovaných korpusech, ale i vztahem synchronie a diachronie v jazyce. S rozvojem synchronních psaných korpusů řady SYN, jejich objemovým nárůstem a vzrůstajícím časovým odstupem mezi nimi se totiž ukázalo, že tyto korpusy v rámci synchronie začínají zachycovat i zárodky diachronie, a že tak nabízejí možnosti pro monitorování jazykového vývoje.

Hlavním cílem práce je popsat možnosti a meze detekce vývojových tendencí v jazyce na materiálu synchronních psaných korpusů řady SYN, které tak tvoří datovou základnu práce. Tato řada obsahuje jak reprezentativní korpusy SYN2000, SYN2005 a SYN2010 zachycující několik blízkých stavů psaného jazyka, tak i rozsáhlé publicistické korpusy SYN2006PUB a SYN2009PUB.

Práce začíná teoretickým úvodem zabývajícím se v kapitole 2 především vztahem mezi korpusem jako vzorkem jazyka a jazykem samým, a v tomto rámci také stavbou korpusů a jejich složením. Hlavním tématem kapitoly 3 je přehled dosavadních studií zaměřených na diachronní srovnání korpusů, odlišnosti jejich přístupů i použitých dat. Žádná z popisovaných studií se však nezabývá přímo diachronním srovnáním blízkých stavů jazyka, které by bylo založené na několika velkých reprezentativních korpusech; obdoba řady SYN2000, SYN2005 a SYN2010 totiž není pro jiné jazyky k dispozici.

Nutnou součástí práce proto bylo hledání cest, jak k problému přistoupit, volba vhodných metod a výběr výchozích dat jako subkorpusů korpusu SYN, který v sobě obsahuje korpusy celé této synchronní řady. Přesto je přehled těchto studií velmi užitečným vodítkem, z něhož vyplývá volba a způsob adaptace metod použitých v hlavní části práce. Vyplývá z něj také důležitost diachronní srovnatelnosti použitých korpusů, a tedy nutnosti použít pro detekci jazykových změn korpusy buď homogenní, nebo naopak reprezentativní a vyvážené podle externích (funkčních) kritérií. V obou případech je důležitým předpokladem konzistentní zpracování a velký rozsah použitých korpusů.

Srovnatelnost v diachronním smyslu je však problematická už z teoretického hlediska kvůli provázanosti jazykových změn se změnami ve společnosti. Ty se projevují například žánrovým posunem v publicistické části korpusu, která je pro detekci vývojových tendencí zásadní. Otevřenou otázkou potom zůstává, do jaké míry by se společenské změny měly odrážet ve složení řady reprezentativních korpusů, nelze-li je dost dobře oddělit od změn jazykových. Jsou-li navíc srovnávaná časová období skutečně blízká, a jsou-li tedy velice blízké i jejich jednotlivé „jazyky“, mohou být vývojové tendence ještě slabé a překryté výraznější synchronní variabilitou.

Prakticky tedy jde o minimalizaci malých nezáměrných rozdílů ve složení srovnávaných korpusů daných pouhým (ne)zařazením konkrétních textů tak, aby rozdíly ve složení korpusů nenarušovaly způsob detekce jazykových změn. Statisticky zjištěnou významnost frekvenčních rozdílů jednotlivých jevů je proto nezbytné na datech zpětně ověřovat a jejich výslednou interpretaci korigovat nejenom intuicí, ale také znalostí přesného složení jednotlivých korpusů.

V kapitole 4 jsou podrobně popsány všechny korpusy řady SYN včetně technických aspektů jejich výstavby, konceptu reprezentativnosti korpusů SYN2000, SYN2005, SYN2010 a způsobu jejich vyvážení založeného na recepci psaného jazyka. Tyto korpusy obsahují psanou češtinu především z let 1990–2009, zachycují v pětiletých intervalech několik blízkých stavů jazyka a zároveň tvoří jednotnou a ucelenou řadu. To platí i přes některé dílčí rozdíly dané nejenom odlišným složením (procentuální zastoupení jednotlivých typů textu a žánrů), ale také zpracováním (různé verze tokenizace, segmentace, lemmatizace a morfologického značkování). Rozdíly dané zpracováním se týkají pochopitelně také publicistických korpusů SYN2006PUB a SYN2009PUB, a proto byly ve všech případech eliminovány aktualizovaným a jednotným zpracováním dat, konkrétně použitím nejnovějších verzí jednotlivých korpusů, které jsou k dispozici jako subkorpusy korpusu SYN (Křen, 2009).

Rozdíly dané odlišným složením lze minimalizovat použitím normalizovaných frekvencí, práce však ukazuje, že je přesto nelze eliminovat úplně. Na základě předchozích dílčích studií, zejména Křen (2007) a Křen a Hlaváčová (2008), byla jako východisko pro diachronní srovnání korpusů zvolena metoda používající statistických měr pro vyhodnocení významnosti rozdílů mezi normalizovanými průměrnými redukovanými frekvencemi (ARF; Savický a Hlaváčová, 2002) pro jednotlivé srovnatelné korpusy, případně jejich subkorpusy. Tato metoda však byla dosud vyzkoušena pouze na lexikální úrovni a pouze při srovnávání dvojic údajů z celých korpusů SYN2000 a SYN2005. Kvůli omezenému rozsahu těchto studií byl navíc důraz kladen na metodu samu, nikoli na popis vývojových tendencí v psané češtině, takže i její vyhodnocení bylo pouze rámcové a další směry výzkumu byly pouze naznačeny.

V kapitole 5 je proto tato metoda popsána v širším kontextu a jsou také zdůvodněna její vylepšení. Jedním z nich je nutnost srovnávání odděleně po hlavních typech textu

v případě reprezentativních korpusů, nebo po jednotlivých letech vydání v případě publicistiky. Vznikly také další, alternativní způsoby, jak zjištěné frekvenční rozdíly statisticky vyhodnocovat. Takto vylepšené metody byly rozšířeny z původní lexikální úrovně (srovnání normalizované ARF lemmat) také na úroveň lexikálních kombinací s aplikací kolokačního filtru.

Volba těchto úrovní je dána především dostupností a spolehlivostí lemmatizace a morfologického značkování synchronních psaných korpusů. Jsme si vědomi omezení vyplývajících z formálního přístupu založeného pouze na lexikonu a jeho kombinatorice, který neumožňuje zkoumat přímo jiné jazykové úrovně jako syntax a zejména sémantiku. V některých případech se k nim však dostáváme nepřímo, prostřednictvím posunů v typických kombinacích a kolokacích, které mohou ukazovat na posuny významu, na významy nové nebo naopak zastarávající.

Metodologicky jde o přístup korpusem řízený (corpus-driven), ovšem s důrazem kladeným na pozdější korigování výsledků a jejich interpretaci také vzhledem k vlivům daným použitými metodami a složením korpusů.

Kapitola 6 představuje v řadě tabulek výsledky použitých metod, a tvoří tak hlavní část práce. Důraz je kladen na podrobné vyhodnocení výsledných tabulek doplněné průběhovými grafy, popisovány jsou zejména příčiny pozorovaného frekvenčního nárůstu, poklesu či oscilace. Ty pak poukazují na obecnější souvislosti nejenom jazykové, ale dotýkající se také věrnosti odrazu psané češtiny ve zdrojových korpusech. Praktickým výstupem této kapitoly je proto souhrn navrhovaných vylepšení v anotaci textů a složení korpusů. Z výsledků jsou kromě proměn doby a tematického zaměření publicistiky patrné také změny v jazyce publicistiky, souhrnně charakterizované jako narůstající neformálnost vyjadřování. Závěrem jsou na základě zjištěných výsledků naznačeny možné obecnější souvislosti a témata vhodná pro podrobnější studium vývojových tendencí některých jevů.

Kromě popsaného hlavního cíle, jímž je zachycení vývojových tendencí v jazyce na korpusech řady SYN, sleduje práce také několik souvisejících cílů vedlejších. Jde především o poskytnutí potřebné zpětné vazby pro další tvorbu synchronních psaných korpusů řady SYN, protože detailní vyhodnocení rozdílů mezi korpusy považovanými za srovnatelné může změnit dosavadní pohled na reprezentativnost v Českém národním korpusu a její význam, a to nejenom v diachronním smyslu. Práce se snaží poukázat na zásadní vliv složení korpusu na jakékoli na něm založené výstupy, a tedy i na nezbytnost správné interpretace korpusových dat. V neposlední řadě si klade za cíl přispět také k osvětlení vztahu mezi korpusem a jazykovou realitou, a korigovat tak dosavadní pohledy. Ty jsou někdy extrémní a pohybují se od odmítání jakékoli výpovědní hodnoty jazykových korpusů až po jejich přeceňování jako zcela nezpochybnitelného zdroje informací o jazyce.

2 Vztah mezi korpusem a jazykovou realitou

2.1 Korpus jako zdroj dat

Dobře známou výhodou korpusové lingvistiky, která se dnes může zdát samozřejmá, je dostatek (někdy až přebytek) empirických dat a také velice snadná kvantifikace „počitatelných“ jevů. Při popisu jazyka tedy není nutné vycházet pouze z intuice a některé přístupy se také snaží se bez ní v maximální možné míře obejít. Intuice je do značné míry subjektivní, může být snadno zavádějící, a proto není vhodné se opírat výhradně o ni. Tím však není zpochybněna její nezastupitelnost při formulování hypotéz a interpretaci výsledků lingvistických výzkumů nebo při zkoumání jevů jen obtížně uchopitelných exaktními metodami, což je také příklad posuzování vztahu mezi korpusem a jazykovou realitou, jak bude vidět v podkapitolách 2.2 a 3.2.

Přístupy k použití korpusu lze rozdělit do dvou skupin, a to na corpus-based a corpus-driven (Tognini-Bonelli, 2001). V dalším textu se budeme držet původních anglických výrazů, protože jejich české překlady se většinou nepoužívají, ačkoli jde v korpusové lingvistice o etablované pojmy. Při ***corpus-based*** (na korpusu založeném) přístupu jde o využití korpusových dat pro pouhé ověřování hypotézy dané předem, tedy formulované intuicí a na základě dosavadních znalostí o jazyce a s využitím tradičních kategorií. Naproti tomu ***corpus-driven*** (korpusem řízený) přístup používání apriorních kategorií a znalostí o jazyce minimalizuje, hlavním (a téměř jediným) zdrojem informací se stává korpus, ze kterého tak mohou vycházet popisy oproštěné od dosavadních, tradičních popisů. Právě minimalizace jejich vlivu umožňuje vznik zcela nových, neotřelých (někdy až extrémních) pohledů na jazyk, což je bezesporu jeho silnou stránkou.

Je však třeba si uvědomit, že ačkoli lze minimalizovat závislost na dosavadních popisech, závislost na konkrétním korpusu se tím naopak výrazně zvyšuje. Každá konkrétní aplikace corpus-driven přístupu a každá statistická metoda totiž stojí na konkrétních datech, žádný jejich výstup nemůže jazykové realitě odpovídat více než korpus sám, jakkoli může být vlastní statistika průkazná a bezchybná. Výsledky těchto metod ani statistická významnost sama o sobě tedy při použití na nevhodný korpus nemusejí vůbec odpovídat významnosti jazykové, takže bez uspokojivého vyřešení vztahu

mezi korpusem a jazykem může tento přístup ve svém důsledku přinášet pouhou iluzi objektivity.

Hlavním tématem této kapitoly je proto právě vztah mezi korpusem a jazykovou realitou, zejména je zdůvodněna potřeba reprezentativnosti a vyváženosti korpusů. Kapitola se dále v přehledu zabývá jazykovými varietami a jejich reprezentací v korpusu, kritérii pro kategorizaci textů i základními druhy korpusů včetně jejich vztahu k homogenitě. Závěrem se zmiňuje také o významu velikosti korpusů, zejména ve srovnání s korpusy internetovými.

2.2 Reprezentativnost korpusů

Prvním problémem, který se váže k reprezentativnosti, je terminologická neujasněnost. V této práci vycházíme z pojmů užívaných Sinclairem (2005), který rozlišuje ***reprezentativnost*** (representativeness) a ***vyváženost*** (balance). Ačkoli sám pojem reprezentativnosti nedefinuje přímo, korpus podle něj reprezentuje jazyk (nebo jeho varietu) tím, že obsahuje texty z něho vybrané. Má-li tedy korpus reprezentovat například psaný jazyk, měl by obsahovat v dostatečném množství vzorky všech psaných variet. Toto tvrzení je velice obecné a každá snaha o jeho praktické naplnění musí vyřešit řadu konkrétních otázek, například jaké množství je dostatečné a jak vybírat vzorky. Hlavními problémy jsou však způsob rozlišování mezi jednotlivými varietami a stanovení poměru, ve kterém by se tyto variety měly do korpusu zahrnovat.

Podle Sinclaira musí být reprezentativní korpus vyvážený tak, aby konkrétní poměry zastoupení jednotlivých druhů textu v něm odpovídaly očekávání: „Roughly, for a corpus to be pronounced balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgements.“ (Sinclair, 2005). Toto vymezení je ovšem vágní v tom, že odkazuje na intuici, a také nijak přesně nevymezuje, co rozumí pod pojmem „druh textu“. Biber et al. (1998, str. 247) naopak tvrdí, že vyváženost není nutná, potřebné je pouze dostatečné pokrytí všech variet, které má korpus reprezentovat. Ani tady se ale nelze vyhnout otázce, v jakém poměru tyto variety do korpusu zařadit. Tento problém bude tématem dalších diskusí, pro začátek pouze shrňme, že korpus reprezentuje určitý jazyk nebo jeho část v jistých proporcích, které by měly být vyvážené. Vztah reprezentativnosti a vyváženosti podobně shrnují i Rayson a Garside (2000, str. 2): „To be representative a corpus should contain samples of all major text types and if possible in some way proportional to their usage in 'every day language'.“

Korpus je pouze více či méně věrným odrazem, vzorkem jazyka nebo některé jeho variety, ale nikdy ho nezastoupí celý. Není to ani nutné, je-li ovšem korpus reprezentativní a odpovídá tak v mnohem menším měřítku celku jazyka (variety). Hlavním problémem, který se k reprezentativnosti váže, je ovšem právě nemožnost objektivně

měřit míru, v jaké korpus jazyku odpovídá. Zdá se přitom zřejmé, že odpověď, kterou nám na tuto otázku může dát intuice, je velice subjektivní a nespolehlivá. Nelze se samozřejmě opírat o intuici založenou na idiolektu jednoho mluvčího, mohla by však reprezentativnost odpovídat „zprůměrované“ intuici, tedy souhrnu mnoha individuálních znalostí jazyka a představ o něm? Lze se spoléhat na to, že by od všech mluvčích zjištěná a „zprůměrovaná“ intuice a odhady typu „které slovo je frekventovanější“ mohly odpovídat korpusu reprezentativnímu například ve smyslu recepce jazyka, když každý jednotlivý mluvčí obtížně odhaduje svou vlastní produkci?

Skeptické jsou v tomto smyslu výsledky studie (Alderson, 2007), která se zabývala problémem spolehlivosti frekvenčních odhadů pokročilých uživatelů jazyka. Autor dochází k závěru, že odhady korpusovým datům neodpovídají, a že se navíc výrazně liší i jednotlivé odhady mezi sebou. Příliš přitom nepomáhá ani jejich zprůměrování, což autor interpretuje jako nenahraditelnost objektivních korpusových dat intuitivními odhady. Na druhé straně však McGee (2008) v reakci na tento článek jeho závěry zpochybňuje a uvádí příklady starších studií, jejichž závěry jsou opačné. Dále kromě řady námitek metodologických poukazuje na to, že i frekvenční data založená na různých korpusech se mohou výrazně lišit, což zpochybňuje i objektivitu korpusových dat. Vysvětluje také, proč nelze očekávat výraznou korelaci mezi frekvenčními odhady jednotlivých mluvčích a korpusovými daty, aniž by mezi nimi nutně musel být rozpor.

Přestože tedy intuice nemusí být úplně špatným vodítkem, zásadním problémem zůstává neexistence jasně definovaných kritérií už pro stanovení toho, co by měl korpus reprezentovat, natož jakým způsobem. Reprezentativnost navíc pochopitelně nelze omezovat jenom na lexikální úroveň, musí zahrnovat také morfologii, syntax a další jazykové roviny. Někteří autoři proto na reprezentativnost rezignují: „The issue of corpus representativeness cannot be usefully taken up unless one specifies the population (in the statistical sense) which we would expect to be faithfully represented. ... So, if we accept the view that there is no agreed standard of comparison, there are serious problems with establishing the criteria for corpus representativeness, and thus the usefulness of the very notion becomes questionable, at least for a general corpus: perhaps specialized corpora or text genres might be more easily dealt with.“ (Lew, 2009, str. 293)

Je však třeba rozlišovat nesnadnost, ne-li nemožnost stanovit objektivní kritéria reprezentativnosti na jedné straně, a jejich naléhavou potřebu na straně druhé. Sinclair (2005) píše: „However unsteady is the notion of representativeness, it is an unavoidable one in corpus design.“ Podobně Biber et al. (1998, str. 246) shrnují „ ... issues of representativeness in corpus design are crucial.“ Na reprezentativnost není možné rezignovat už proto, že bez ní nelze korpusová data interpretovat ve vztahu k jazyku. Musíme se však smířit s tím, že tento vztah není snadno kvantifikovatelný, korekce intuicí se zdá být nezbytná, i když jde samozřejmě o subjektivní prvek. Leech (1991,

str. 27) dokonce píše: „We should always bear in mind that the assumption of representativeness must be regarded largely as an act of faith.“ Tuto větu bychom si však měli vykládat jako střízlivé konstatování skutečnosti, týž autor (Leech, 2004, str. 70) uvádí pět předpokladů potřebných k přechodu od popisu dat k popisu jazyka, tedy k extrapolaci výsledků získaných studiem korpusových dat na jazyk jako celek nebo na některou jeho varietu, a to konkrétně při diachronním srovnávání brownovských korpusů (tyto korpusy jsou podrobněji popsány v oddílu 3.3.4). Za největší problém přitom považuje právě vztah korpusu a jazyka, konkrétně otázku, zda je korpus dost velký a vyvážený, tedy zda daný jazyk nebo jeho varietu skutečně reprezentuje. Zmiňuje také potřebnou srovnatelnost korpusů i to, že statisticky signifikantní výsledky nemusejí být způsobeny jenom jazykovými rozdíly. Dodává však: „In my view, none of these hazards justifies a response of extreme scepticism which says if one cannot prove the truth of these descriptions, one should not make them at all.“ (Leech, 2004, str. 71)

2.3 Jazyk a jeho variety

Přirozený jazyk je mnohorozměrný a složitě vnitřně provázaný fenomén, navíc velice variabilní, který je obtížné, ne-li nemožné uchopit jako celek. Už samotnou volbou úhlu pohledu ho nutně redukujeme a zplošťujeme, a to i v případě, že se při popisu omezíme pouze na jednu jeho funkčně, regionálně a časově podmíněnou varietu. Tato podkapitola si klade za cíl zdůraznit vzájemnou odlišnost některých variet, což dále zvyšuje důležitost vyváženosti korpusů: „One particularly important concern for corpus design is diversity. Throughout this book, the analyses show that there are important differences in the use of lexical, grammatical, and discourse features across different varieties of language. In fact, the analyses show that there is really no such thing as 'general language'; each register has its own patterns of use. Thus, any corpus that is used for studies of variation or that seeks to represent a language needs to be concerned with the diversity of texts that it includes.“ (Biber et al., 1998, str. 248)

Terminologie používaná u jazykových variet je ještě neustálenější než v případě reprezentativnosti. V anglické literatuře jsou pro ně běžně užívány pojmy domain, register, sublanguage, text type, genre, topic nebo style bez jasného a obecně přijímaného vymezení rozdílů mezi nimi (Lee, 2001). Přestože jsou rozdíly mezi některými z nich (např. topic a style) zřejmé, není už jasný jejich vztah k dalším podobným pojmům a jejich případné překrývání. Někteří autoři zavádějí užitečné distinkce, které se ovšem v daném smyslu neužívají obecně, například Biber (1988) používá pojem genre pro varietu definovanou na základě externích kritérií, zatímco text type pro varietu definovanou na základě kritérií interních (o tomto tématu podrobněji dále); je přitom samozřejmě možné, že dva texty stejného žánru mohou mít různý typ textu a naopak. Podrobný rozbor všech uvedených pojmů a jejich chápání jednotlivými autory by však

nebyl pro cíl této práce účelný, v dalším textu se proto budeme držet zastřešujícího pojmu ***varieta*** a rozlišovat *typ textu* a *žánr* ve smyslu používaném v ČNK (podrobněji dále v oddílu 4.2.4 a v podkapitole 4.3). Výjimkou z uvedené zásady budou pouze případy dočasného převzetí pojmů používaných v citované literatuře.

Jak ukazují Křen a Hlaváčová (2008, str. 443) na příkladu předložky *v*, i velice frekventovaná funkční slova mohou mít mezi hlavními typy textu nerovnoměrnou distribuci. Mnoho dalších příkladů lze najít také ve *Statistikách češtiny* (Bartoň et al., 2009) nebo ve *Frekvenčním slovníku češtiny* (Čermák, Křen et al., 2004), v němž jsou frekvence všech hesel udávány společně s jejich rozložením ve třech hlavních typech textu. Slovník uvádí celkem 423 hesel z celkových 50 000, jejichž výskyt v některém typu textu převažuje natolik výrazně, že se ve druhých dvou téměř nevyskytuje (udávané normalizované procento výskytů daného hesla v každém z nich je po zaokrouhlení 0 %). Také ve slovnících řady *Routledge Frequency Dictionaries* se stalo běžnou praxí uvádění typického typu textu pro každé heslo. Ve *Frequency Dictionary of Czech* (Čermák, Křen et al., 2011) je téměř polovina všech hesel (konkrétně 2 346 z celkových 5 000) označená jako nerovnoměrně rozložená ve zdrojových korpusech.

Významný vliv typu textu se samozřejmě neomezuje jen na češtinu a na úroveň lexémů. Johansson a Hofland (1989) ve své obsáhlé studii ukazují na korpusu LOB mj. výraznou závislost frekvencí jednotlivých morfosyntaktických značek (tagů) na „žánru“ textu v angličtině. Podobně Mair et al. (2002) popisují dílčí studii zabývající se srovnáním frekvencí tagů mezi brownovskými korpusy LOB a FLOB (viz oddíl 3.3.4), a to jak na celých korpusech, tak na jejich čtyřech subkorpusech (press, general prose, learned (odborné texty) a fiction) vzniklých seskupením původních 15 textových kategorií; její výsledky ukazují, že frekvence jednotlivých jevů jsou na nich výrazně závislé. Také na korpusu založená práce *Longman Grammar of Spoken and Written English* (Biber et al., 1999) rozlišuje čtyři situačně definované „registers“ (conversation, fiction, news a academic prose) a uvádí u jednotlivých jazykových jevů (ne)rovnoměrnost distribuce mezi nimi jako důležitou součást popisu těchto jevů. Je vidět, jak se rozložení třeba i jen frekvence jednotlivých slovních druhů dramaticky liší (např. normalizovaná frekvence substantiv v publicistice je více než dvojnásobná oproti mluvenému jazyku); obdobné rozdíly lze přitom najít u mnoha dalších jevů na různých jazykových rovinách.

Známým způsobem studia jazykové variability je vícerozměrná analýza (multidimensional analysis; Biber, 1988; Biber, 1995). Základem tohoto přístupu je stanovení relevantních jazykových rysů (linguistic features) a kvantifikace jejich výskytu v analyzovaných textech. Tyto rysy mohou být poměrně různorodé, důležitá je možnost jejich spolehlivé identifikace v textech. Může jít například o zájmena 1. a 2. osoby, ukazovací zájmena, modální slovesa, pasivum, druhy vedlejších vět, celkový počet substantiv nebo průměrnou délku slova. Dalším krokem je seskupení těchto rysů do skupin podle jejich společného výskytu v analyzovaných textech (skupin bývá obvykle 5 až 10);

to je výsledkem faktorové analýzy, což je standardní statistická metoda. Může se například ukázat, že vysoká frekvence výskytu zájmen 1. a 2. osoby v analyzovaných textech výrazně koreluje s vysokou frekvencí ukazovacích zájmen, malou frekvencí pasiva, malým počtem substantiv a kratšími slovy. Všechny tyto rysy tedy utvoří jednu skupinu souvisejících rysů, do které však nejsou zařazeny rysy, které s nimi dostatečně silně nekorelují; příkladem by mohla být frekvence užívání modálních sloves.

Vzniklé skupiny rysů jsou interpretovány jako variační dimenze (dimensions of variation), které sice nejsou v textech přímo měřitelné, ale implicitně určují frekvence daných jazykových rysů. Tato interpretace je zpravidla funkční, ve výše uvedeném případě by mohlo jít o dimenzi vyjadřující míru interaktivity (involved vs. informational). Nakonec se pro každý analyzovaný text spočítají tzv. dimension scores, tedy hodnoty určující, kde se text nachází na škále dané každou z dimenzí. Tyto hodnoty se pro každý text v daném registru zprůměrují a tím se získá pozice každého registru na škále v každé dimenzi. Vícerozměrná analýza tak může být použita jako teoreticky a kvantitativně jasně vymezený způsob definice řady variet potvrzující například fakt, že jazyk beletrie je v mnoha ohledech jiný než jazyk odborné literatury nebo že v přepisech rozhlasových pořadů najdeme jiný typ jazyka než v běžné neformální konverzaci. Jde tedy – s výjimkou počátečního stanovení jazykových rysů – o corpus-driven přístup ke kategorizaci textů, který umožňuje také sledování variability a vývoje takto vzniklých skupin.

Použití vícerozměrné analýzy pro diachronní studie ukazují Biber a Finegan (2001), její použití při srovnávání angličtiny s typologicky i kulturně odlišnými jazyky (korejštinou, somálštinou a tuvalštinou) potom Biber (1995). Výsledky těchto studií jsou popsány také v oddílu 3.3.2, v tuto chvíli je podstatné, že při srovnávání variačních dimenzí a registrů pro každý z těchto jazyků Biber dochází k závěru podporujícímu hypotézu o jejich univerzalitě: „The present study shows that even when registers are defined at a high level of generality (e.g., conversation, editorials, personal letters), and even when comparisons are across markedly different language families and cultures, parallel registers are indeed more similar cross-linguistically than are disparate registers within a single language.“ (Biber, 1995, str. 279)

Zdůrazňovaná různost jednotlivých jazykových variet souvisí také se vztahem mezi jazykovým centrem a periferií. Zatímco centrum je typicky složené z frekventovaných, relativně bezpříznakových jednotek s množstvím významů a funkcí, periferie je opakem centra, přičemž můžeme uvažovat ještě poměrně široké přechodové pásmo. Jak uvádí Čermák (v tisku), problémem je absence spolehlivých kritérií pro odlišení centra a periferie, která souvisí i s mnohorozměrností tohoto jevu. Jde tady nejenom o polysémii (centrální význam může být doplněn periferními) nebo o vztah k jazykovému vývoji (jednotky se přesouvají z centra do periferie a naopak), ale také o závislost všech těchto jevů na konkrétní varietě. Zjednodušeně se dá říci, že zatímco jednotky z centra jazyka

se typicky vyskytují ve všech jeho varietách, směrem k periferii se závislost na varietě výrazně zvyšuje: co se běžně používá v jedné, může být ve druhé na hranici potenciality. Jinými slovy, typičnost je do značné míry daná právě varietou a celojazykového je relativně málo. Je přitom potřeba zdůraznit, že tento mnohorozměrný prostor je škálovitý ve všech směrech, nejenom na ose centrum – periferie, ale i při přechodu mezi jednotlivými stavy jazyka atd.

Chceme-li však s jazykem pracovat a popisovat ho, je přiměřené zjednodušení této jeho inherentní složitosti nevyhnutelné. Soustředěním na určitý jev by však neměly unikat zákonitosti obecnější povahy, proto je nezbytné si omezení daná námi zvoleným úhlem pohledu neustále uvědomovat a brát je v úvahu při volbě pracovního postupu i interpretaci výsledků. Je-li popis jazyka založený na korpusu jako jeho vzorku a má-li být při interpretaci výsledků možné vztáhnout popis korpusu na celý jazyk, je nutné se vážně zabývat i vztahem jazyka a korpusu jako důležitého mezistupně mezi jazykem a jeho na korpusu založeným popisem. Vracíme se tak k vytváření korpusů, které by měly jazyk (případně některou z jeho variet) jako vzorek reprezentovat, ke složení tohoto vzorku, k jeho vyváženosti a k otázce, zda by se daný korpus měl snažit být obecný, celojazykový, nebo „pouze“ úžeji specializovaný, jak bude podrobněji popsáno dále.

2.4 Kritéria pro kategorizaci textů v korpusech

Aby bylo možné vytvořit reprezentativní korpus s vyváženým podílem všech v něm zastoupených variet, je nejdřív nutné vyřešit problém kategorizace textů ve vztahu k těmto varietám. Existují dva odlišné, ortogonální přístupy pro kategorizaci textů založené buď na externích, nebo interních kritériích. Zatímco ***externí kritéria*** jsou předem daná, tradiční mimotextová kritéria vztahující se ke komunikační funkci textu a beroucí v úvahu její účel, publikum atd., ***interní kritéria*** berou v úvahu pouze frekvenci a vzájemné vztahy nejrůznějších lexikálních a gramatických jevů v textu beze vztahu ke kritériím externím (příkladem takové klasifikace může být vícerozměrná analýza popsaná v předchozí podkapitole).

Pokud je nám známo, jsou texty ve všech velkých světových korpusech kategorizovány na základě externích kritérií, což má několik důvodů. V první řadě jde o kritéria tradiční a uživateli očekávaná, používající kategorie, jako jsou (na různé úrovni podrobnosti) například beletrie, román nebo historický román. Neméně důležitý je fakt, že se s nimi snadněji pracuje: kategorizaci lze provádět víceméně intuitivně, pouze s pomocí základních instrukcí řešících sporné a nejednoznačné případy (např. rozdíl mezi románem a povídkou; jak kategorizovat humoristický historický román z vojenského prostředí apod.). Používání externích kritérií má na druhou stranu několik závažných nevýhod. Je zřejmé, že tato kategorizace je subjektivní a že může být obtížné udržet

konzistentní anotaci i v rámci jednoho projektu. Každý z projektů navíc používá jiné rozlišení na jiném teoretickém základě, takže například kategorizace textů v ČNK neodpovídá BNC (British National Corpus) apod. Snadnost aplikace externích kritérií navíc svádí ke zjednodušování problému, zvláště při používání příliš vágních a ad hoc definovaných kategorií. Meurman-Solin (2001, str. 13) o tomto problému a jeho možných důsledcích píše: „ ... there seems to be a risk that the information coded into corpora is uncritically accepted as a frame of reference in the analysis and discussion of the findings.“

Na druhou stranu je konzistentní používání dobře definovaných externích kritérií vhodné z teoretického hlediska. Má-li korpus reprezentovat daný jazyk nebo varietu, je potřeba při výběru textů vycházet především z různých komunikačních situací a kontextů a teprve z takto sestaveného korpusu může vyplynout vztah mezi nimi a interními jazykovými rysy. Tyto jazykové rysy tak nejsou na kritériích, podle nichž byl korpus sestaven, závislé přímo, i když nepřímé souvislosti tady samozřejmě jsou. Někteří autoři proto kategoricky tvrdí, že korpusy by se měly vytvářet pouze na základě externích kritérií, například Sinclair (2005) píše: „The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise. ... Selection criteria that are derived from an examination of the communicative function of a text are called *external criteria*, and those that reflect details of the language of the text are called *internal criteria*. Corpora should be designed and constructed exclusively on external criteria.“ Podobně argumentují Králík a Šulc (2005, str. 358–359) pohledem z hlediska uživatele, jeho očekávání a univerzality možných použití korpusu. Také *Longman Grammar of Spoken and Written English* (Biber et al., 1999) uvádí normalizované (a tedy srovnatelné) frekvence jednotlivých jazykových jevů pro čtyři situačně (tj. externě, nikoli interně) definované „registers“.

Interní kritéria jsou naproti tomu objektivně kvantifikovatelná, i když někdy obtížněji uchopitelná (srov. názvy pro typy textu, které jsou výsledkem Biberovy vícerozměrné analýzy). Sigley (1997, str. 231–232) proto argumentuje proti kategorizaci textů na základě předem daných externích kategorií takto: „A major feature of the approach used here (as of Biber's) is that it allows researchers to question the text categories used in collecting the data, and if necessary to re-classify texts according to criteria important for their study. ... Corpus analysts are recommended to beware of treating the pre-existing text categories as natural groups, and to consider alternative text groupings which may be more relevant for their purpose.“

Kdybychom se však při tvorbě korpusů opírali pouze o interní kritéria, zůstane otázka, jaký je vztah na nich založených kategorií k různým komunikačním situacím, a hlavně jakým způsobem tyto kategorie v korpusu vyvážit: „The initial selection of texts for inclusion in a corpus will inevitably be based on external evidence primarily . . .

A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation.“ (Atkins et al., 1992, str. 5). Ačkoli se tedy při vytváření korpusů nelze vyhnout používání externích kritérií jako hlavního způsobu kategorizace textů, je jistě vhodné používat interní kritéria jako zpětnou vazbu pro studium textů předem externě kategorizovaných, pro analýzu vztahu mezi konkrétními jazykovými jevy a komunikační funkcí jednotlivých variet, a také k případným korekcím složení korpusu ještě před jeho zveřejněním. Jak poznamenává Biber (1993), při výstavbě korpusů je nutné se soustředit na pokrytí všech situačních (tj. externě definovaných) parametrů už proto, že nelze nijak předem definovat kategorie založené na interních kritériích, a dodává, že „ ... the results of previous research studies, as well as on-going research during the construction of a corpus, can be used to assure that the selection of texts is linguistically as well as situationally representative.“ Biber (1993, str. 245). Toto použití interních kritérií pro případnou pozdější korekci původní externí klasifikace se tedy zdá být ideálním styčným bodem obou přístupů, přestože konkrétní provedení a rozsah této korekce zůstává otevřenou otázkou.

2.5 Druhy korpusů a homogenita

Jazykové korpusy můžeme z hlediska šíře popisu jazyka rozdělit na obecné a specializované. Zatímco ***obecný*** korpus se snaží postihnout jazyk jako celek, případně jeho velkou část (např. korpus psaného jazyka), korpus ***specializovaný*** naproti tomu jenom oblast úzce vymezenou (např. korpus odborných lingvistických textů). Tato dichotomie má samozřejmě poměrně široké přechodové pásmo, ukazuje však na souvislost jednak s reprezentativností, jednak s homogenitou korpusů. Obecné korpusy jsou už z definice heterogenní, a mají-li navíc jazyk skutečně reprezentovat, musejí být i vyvážené, zatímco korpusy specializované jsou relativně homogenní, pojetí reprezentativnosti se u nich zjednodušuje a také vyváženost přestává být zásadním problémem.

Prozatím jsme při dosavadním popisu problémů s reprezentativností korpusů a odlišováním jednotlivých jazykových variet implicitně předpokládali korpusy obecné, u kterých je situace nejsložitější. Příkladem jejího zjednodušení u specializovaného, homogenního korpusu může být ORAL2008, sociolingvisticky vyvážený korpus neformální mluvené češtiny (Waclawičová et al., 2009). Do tohoto korpusu byly zařazeny pouze přepisy nahrávek soukromé mluvené komunikace, jejíž účastníci byli fyzicky přítomni, dobře se znali a mluvili na témata, která nebyla nijak předem daná; typicky šlo o rozhovory v rodině nebo mezi přáteli. Pro jeho vyvážení byla proto použita kritéria sociolingvistická, konkrétně pohlaví, věk, vzdělání a místo pobytu mluvčích v dětství; přesný postup vyvažování popisují Waclawičová a Křen (2008, str. 110–112). Tato kritéria jsou jednoznačně externí, v tomto případě si navíc lze jen obtížně představit smysluplné zapojení interních kritérií.

Velké reprezentativní korpusy, které se snaží obsáhnout jazyk jako celek, naopak homogenní záměrně nejsou: „Large corpora should not really be thought of as homogeneous wholes, but rather as sets of overlapping subcorpora. For human language in use is very domain-specific.“ (Hanks, 2000, str. 9) Snaží se sice být reprezentativní v tom smyslu, že by se v nich měly vyskytovat jevy typické pro všechny zastoupené jazykové variety, jejich konkrétní podíly na složení takových korpusů ale mohou mít na výstupy zásadní vliv. Vezmeme-li navíc v úvahu různorodost těchto korpusů, je otázka, zda je vůbec smysluplné snažit se popsat jazyk jako celek. Homogenita korpusů je navíc vhodná i pro statistické popisy, i když je samozřejmě pravda, že přesné vymezení určité variety není vůbec jednoduché ani jednoznačné.

Na tomto místě bychom chtěli poukázat také na vztah mezi měřením homogenity a srovnáváním korpusů. Kilgarriff (2001) se snaží rozdíl mezi korpusy kvantifikovat a ukazuje, že měření homogenity korpusu lze převést na řadu srovnání náhodně vybraných dvojic jeho subkorpusů a že lze pro tyto účely použít i stejnou míru. Platí totiž, že čím je korpus heterogennější, tím jsou jeho subkorpusy odpovídající jednotlivým jazykovým varietám odlišnější, a to i z hlediska statistického popisu. Při snaze o generalizaci těchto parciálních popisů na korpus jako celek (jejich „zprůměrování“) však mezi nimi dochází spíše k interferencím, celkový výstup je pak jen obtížně interpretovatelný a o jazyce toho příliš nevypovídá, a to ani jsou-li data v korpusu k dispozici v dostatečném množství: „Extrémní růst počtu sledovaných jevů (rozsah textů) totiž nemusí – v lingvistice zejména – nutně znamenat extrémní zpřesnění poznatků. Nepohybujeme se v homogenním prostoru. ... Korpus je proto třeba vidět spíše jako soubor odhadů, než jako celistvý prostor, z něhož by mohla plynout souhrnná interpretovatelná hodnota nějaké kvantitativní charakteristiky. V tomto pohledu korpus nepředstavuje a ani nemůže představovat jazyk jako celek, pro žádný jev nemůže zastupovat jedinou, univerzální distribuci, ale vždy jen součet mnoha.“ (Králík, 2006, str. 207) A dále: „Tento závěr se shoduje se sílícím doporučením zkoumat spíše homogennější jazyk jednoho typu, např. žánru, než celý korpus.“ (Králík, 2006, str. 209)

Jak již bylo zmíněno výše, Sinclair (2005) pro výběr textů do korpusu jednoznačně doporučuje používání externích kritérií. Zdůrazňuje však také výhody nezařazování textů příliš netypických, které jsou „unrepresentative of the variety“; preferuje tak vznik homogennějších subkorpusů uvnitř obecného korpusu: „There is a balance to be struck between coverage and homogeneity in the attempt to achieve representativeness. ... A corpus should aim for homogeneity in its components while maintaining adequate coverage.“ (Sinclair, 2005) Jde tedy o pohled na obecný korpus jako na strukturovaný soubor mnoha jasně odlišených specializovaných, homogenních subkorpusů obsahujících pouze texty typické pro danou specifickou varietu, nikoli pouhý souhrn textů bez přesného určení variety a s mnoha překryvy mezi nimi. V tomto smyslu se mu také blíží pojetí monitorovacích korpusů popisovaných dále v oddílu 3.3.5, kde jde

navíc i o udržení konzistence těchto subkorpusů v čase. V této souvislosti bychom chtěli upozornit, že právě čas je faktor způsobující vzrůstající nehomogenitu na první pohled úzce vymezených korpusů, jako je například korpus vytvořený pouze z týdeníku TIME (Millar, 2009, str. 214) nebo MF Dnes (Křen a Hlaváčová, 2008, str. 444–445); tento problém probíráme podrobněji v podkapitole 3.2.

2.6 Velikost korpusu a volba vzorku

Potřebná velikost korpusu je další důležitou otázkou, odpověď na ni se však od počátků korpusové lingvistiky měnila. V zásadě platilo, že maximální dostupná velikost korpusu je dostatečná, ať už jí byl 1 milion slov korpusu Brown na začátku 60. let, 100 milionů BNC o třicet let později nebo současné miliardy ve webových korpusech, IDS Mannheim nebo ČNK. V současné době je ovšem shoda v tom, že pro dostatečné zastoupení různých jazykových rovin (případně konkrétních jazykových jevů) je potřeba různá velikost korpusu. Obecně platí, že stačí menší korpus pro jevy frekventované (např. gramatické), naopak velký korpus je potřeba pro jevy méně časté (např. lexikální). Přestože zůstává otázkou, zda může být vůbec dosaženo takové velikosti korpusu, která by byla dostačující univerzálně, tedy pro všechna možná použití, je vždy lepší mít k dispozici více dat i v případě, kdy to není nezbytné.

V důsledku tohoto faktu, problémů s reprezentativností korpusu, a zároveň snadné přístupnosti velkého množství elektronických textů, se objevil zjednodušující slogan „the more the better“, který vyjadřuje víru, že ve velkých datech se vše ztratí (nebo naopak najde). Je samozřejmě pravda, že čím větší korpus je k dispozici, tím spolehlivější a statisticky průkaznější jsou také veškeré na něm založené údaje. Na druhé straně je však zřejmé, že přidáváním textů stále stejného druhu (např. novinových článků do publicistického korpusu) nelze dosáhnout dostatečného zastoupení rysů typických pro jinou varietu. Jinými slovy nelze podceňovat vztah mezi korpusem a jazykovou realitou, kterou korpus reprezentuje. Může se pak snadno stát, že výsledky založené na nereprezentativním korpusu budou platit jen pro korpus sám, o jazyce přitom nemusejí vypovídat téměř nic, podobně jako nikoho nepřekvapí, že fakta získaná analýzou korpusu složeného výhradně z publicistiky neplatí pro popis mluveného jazyka. Velikost korpusu sama o sobě tedy nemůže nahradit jeho vhodné složení: „However, for all kinds of research, it is important to realize that size cannot make up for a lack of diversity.“ (Biber et al., 1998, str. 249)

Volba vzorku byla aktuální v době, kdy byly korpusy malé, počítače pomalé a úložný prostor drahý. Tehdy bylo – zejména z kapacitních důvodů – praktické zařazovat do korpusu pouze části textů, čímž ovšem vznikala otázka, jak velké vzorky z textů vybírat a jakým způsobem. Mezi korpusy složené ze vzorků patří korpusy Brown, LOB, Helsinki Corpus nebo ARCHER, popsané podrobněji v podkapitole 3.3. Při výběru vzorku

z textu však vzniká problém se ztrátou kontinuity v místě řezu a také s tím, kterou část textu vybrat; jeho jednotlivé části (úvodní, závěrečná atd.) se totiž mezi sebou typicky liší, jak ukazují například Biber et al. (1998, str. 166–169). V současné době už není důvod zařazovat do korpusu jenom část textu, s výjimkou řídkých případů, kdy je text obtížně získatelný (např. přepisováním z rukopisu), není k dispozici celý nebo kdy je příliš rozsáhlý vzhledem ke kategorii, kterou by měl v korpusu reprezentovat.

V této souvislosti ještě stručně zmiňme různost přístupů při výběru textů pro reprezentaci dané kategorie (např. historický román) v korpusu. Zatímco u korpusů Brown, LOB nebo ARCHER šlo o částečně náhodný výběr textů z bibliografických registrů ještě před zahájením jejich zpracování, výběr textů pro Helsinki Corpus nebo reprezentativní korpusy řady SYN v ČNK byl do jisté míry záměrný a sledoval několik často protichůdných kritérií; ta jsou uvedena v oddílu 4.2.6.

2.7 Tradiční jazykový korpus a internet

Budování jazykových korpusů je v posledních letech stále snadnější nejenom proto, že téměř všechny texty dnes vznikají na počítači, ale především díky nástupu webu a s ním spojeným možnostem přístupu k obrovskému množství textů v elektronické podobě. Je proto přirozené, že se záhy začaly objevovat snahy o využití webu pro lingvistické účely.

Jedním ze způsobů, který se objevil poměrně brzy, je vývoj a používání speciálních nadstaveb nad populárními vyhledávači (AltaVista, Google), které tyto vyhledávače používají přímo pro hledání on-line, ale které zároveň výsledky vyhledávání dále filtrují a zpracovávají tak, aby připomínaly tradiční konkordanční řádky. Příklady těchto nástrojů mohou být KWiCFinder (Fletcher, 2004) nebo WebCorp (Renouf et al., 2005). Tento způsob využívání webových dat má však několik zásadních nevýhod: především je plně závislý na způsobu fungování těchto vyhledávačů, od formátu poskytovaných dat a jejich spolehlivosti po omezené možnosti dotazovacího jazyka a způsobu třídění (ranking) výsledků; vše je přitom podřízeno primárně vyhledávání informací, nikoli jazykových jevů. Často je také nemožné zjistit údaje o vzniku textu, autorství apod., zatímco tradiční korpusy jsou většinou nejenom bibliograficky anotované, ale mohou být také strukturované (označení odstavců a vět), lemmatizované a morfosyntakticky označkované. Dotazovací jazyk korpusových manažerů navíc bývá velice bohatý a umí tuto anotaci využívat právě pro lingvisticky orientované dotazy, statistiky kolokací apod. Jak dále ukazuje například Davies (2011), je kromě nespolehlivosti metadat velkým rizikem webových vyhledávačů také nespolehlivost jimi uváděných frekvencí. V neposlední řadě je zásadní nevýhodou tohoto přístupu neopakovatelnost: týž dotaz položený v různém čase může dávat jiné výsledky, což je při vědecké práci problémem samo o sobě. Navíc i příčiny těchto rozdílů mohou být různé, od změn webu jako stále aktualizovaného média po nečekané (a nijak nezdokumentované) změny algoritmů vy-

hledávacích služeb a jejich přizpůsobování konkrétnímu uživateli, což téměř znemožňuje jakoukoli interpretaci zjištěných rozdílů.

Vhodnějším způsobem využití webových dat je proto tvorba webových (web-crawled) korpusů (Baroni et al., 2006; Sharoff, 2006; Baroni et al., 2009). ***Webové korpusy*** jsou rozsáhlé korpusy o velikosti v řádech miliard slov vzniklé automatickým stažením relevantních textů z webu (pro konkrétní jazyk určených zpravidla klíčovými slovy) a jejich následným zpracováním nástroji, které zajišťují plně automatickou konverzi, čištění, detekci (semi-)duplikátů, lemmatizaci, morfologické značkování, žánrovou klasifikaci atd. Jde tedy o rychlou a zjednodušenou variantu k tradičnímu budování korpusu, takto připravené korpusy je přitom možné používat stejným způsobem jako korpusy tradiční, čímž padá většina námitek zmíněných v předchozím odstavci. Webové korpusy jsou navíc dostatečně velké, aktuální, jejich tvorba je mnohem snadnější, rychlejší (a tedy i levnější) a také parametrizovatelná podle zamýšleného způsobu použití.

Na druhé straně je třeba poukázat na několik nevýhod takového přístupu: těmto korpusům chybí podrobná bibliografická anotace (uvedení autora, roku vydání atd.), automatická žánrová klasifikace je navíc nespolehlivá a musí se opírat výhradně o interní kritéria (viz podkapitola 2.4; to však lze považovat i za výhodu). Automatické čisticí procedury jsou obvykle použité jednotným způsobem na celá data a nemají kvalitu a přesnost cílených poloautomatických metod používaných například v ČNK (Křen, 2009). Konečně nejzávažnější námitka se týká složení takto získaných korpusů: zatímco webové korpusy typicky obsahují množství nových, ryze internetových žánrů, chybějí jim naopak v dostatečném množství klasické žánry, kterých je na webu málo, zejména beletrie. Tento rozdíl ve složení tradičních korpusů na jedné straně a korpusů webových na straně druhé nelze považovat za výhodu ani nevýhodu webových korpusů, je však nutné ho vzít v úvahu před každým použitím takto získaných dat. Webový korpus může být díky své velikosti a dalším výše zmíněným výhodám pro mnoho aplikací vhodnější volbou, přesto je problematické a ničím nepodložené a priori prohlásit velký webový korpus za lepší vzorek psaného jazyka, než je korpus tradiční. Dostáváme se tady opět k potřebě reprezentativnosti a vyváženosti jazykových korpusů, kterou – jak jsme ukázali v podkapitole 2.6 – nelze nahradit prostým množstvím dat.

3 Srovnávání korpusů

3.1 Synchronní a diachronní srovnání

Korpusy lze srovnávat mnoha způsoby a na jejich srovnání pohlížet z několika hledisek. Pro účely této práce je nejdůležitější rozdíl mezi synchronním a diachronním srovnáním korpusů. Za ***synchronní*** považujeme takové srovnání, při němž se všechny rozdíly nalezené mezi srovnávanými korpusy považují za známky jejich rozdílného složení. Příklady několika takových přístupů založených většinou na frekvenčních seznamech tvarů nebo lemmat popisují Rayson a Garside (2000) nebo Kilgarriff (2001), zmíněný v podkapitole 2.5 týkající se homogenity korpusů. Je přitom možné navzájem srovnávat jak dva přibližně stejně velké korpusy, tak i malý korpus vzhledem k velkému referenčnímu korpusu. Zjištěné rozdíly není nutné ani nijak interpretovat, v některých případech stačí jejich prostá kvantifikace, tj. stanovení míry shodnosti (rozdílnosti). Jak ovšem zdůrazňují Rayson a Garside (2000), pro smysluplné srovnání je potřeba vzít v úvahu řadu dalších kritérií, zejména reprezentativnost, vnitřní homogenitu, vzájemnou srovnatelnost korpusů a také spolehlivost statistických testů.

Pro ***diachronní*** srovnání je však na rozdíl od synchronního podstatné odlišovat rozdíly způsobené jiným složením korpusů od rozdílů způsobených jazykovým vývojem. Synchronní srovnání je tedy jednodušší v tom smyslu, že jakýkoli zjištěný rozdíl je považovaný za relevantní, zatímco cílem diachronního srovnání je detekce jazykových změn. Znamená to, že se předpokládá jistá míra rozdílnosti korpusů, avšak tyto rozdíly by měly být v ideálním případě způsobené výhradně jazykovým vývojem, nikoli odlišným složením korpusů nebo výběrem textů (Křen a Hlaváčová, 2008, str. 483).

Korpusy, které by se lišily pouze na časové ose, však v praxi neexistují. Toto tvrzení platí pochopitelně obecněji, protože variabilita je v jazyce inherentní, mnohorozměrný jev, vlivy jednotlivých faktorů se střetávají a je téměř nemožné je od sebe oddělit, což platí i při synchronním pohledu. Proto je také obtížné určit konkrétní příčinu jednotlivých rozdílů zjištěných při (nejenom) diachronním srovnání jednotlivých korpusů a tyto rozdíly interpretovat: „ ... each of these sources of linguistic variation can obscure any of the others. ... When carrying out comparisons, however, it is often difficult to build corpora that only differ on a single factor.“ (Oakes, 2009, str. 182)

Přes právě zmíněná praktická úskalí je však diachronní srovnání korpusů velice užitečné a lákavé i proto, že se při něm ze synchronního pohledu obtížně uchopitelná variabilita proměňuje ve zřetelnější vývojové tendence. Objevilo se proto množství studií, které se diachronním srovnáním zabývají, a právě jejich přehled je spolu s problémem reprezentativnosti v diachronním smyslu hlavním tématem této kapitoly.

3.2 Reprezentativnost diachronních korpusů

Při návrhu a vytváření diachronních korpusů je potřeba zajistit reprezentativnost na další, časové ose, což je zvláště obtížné v případě obecných korpusů mapujících vývoj celého jazyka po mnoho století. Vzhledem k časovému odstupu je také mnohem obtížnější (ne-li nemožné) stanovit v daném období proporce vyváženosti jednotlivých variet, protože není možné se opřít o jazykové povědomí uživatelů jazyka, a kromě toho přibývá i praktický problém (ne)dostupnosti textů konkrétního typu z určitého časového období (Kučera, 2011, str. 66–67). Reprezentativnost v diachronním smyslu je tedy ještě problematičtější než ve smyslu synchronním, protože se objevují další obtížně měřitelné veličiny a s tím spojené metodologické problémy. Jedním z nich je také provázanost jazykových změn se změnami společenskými a kulturními a míra, v jaké by to měl korpus odrážet. Tato provázanost se projevuje mj. tím, že některé variety vznikají (např. blogy), vliv jiných na jazyk dané doby se naopak výrazně snižuje (náboženské traktáty).

Dokonce i v případě, kdy je zastoupení dané variety zdánlivě stabilní, může docházet k podstatným změnám zevnitř. Pokud je navíc časový rozdíl mezi jednotlivými texty náležícími do „stejné“ variety příliš velký, může jít ze synchronního pohledu o varietu jinou. Biber et al. (1998, str. 210–215) uvádějí jako příklad anglické odborné lékařské texty, kterými byly na začátku 18. století často případové studie psané formou osobních dopisů, zatímco pro konec 20. století je typický velice hutný a neosobní odborný styl; podobně výrazný posun, i když jiným směrem, lze vysledovat i u dramat. Přestože jsou oba tyto extrémy ze synchronního pohledu naprosto odlišnými typy textu (v Biberově smyslu), bylo při vytváření korpusů Helsinki Corpus a ARCHER (viz oddíl 3.3.2) rozhodnuto ponechat vše v témže „registru“ (medical prose, resp. drama). I k těmto výrazným posunům se tedy přistupuje jako ke změně uvnitř dané variety, což nijak nemění její celkové zastoupení v jednotlivých časových obdobích. Tady se opět dostáváme k různým způsobům vymezení jednotlivých variet, který jsme již probírali v podkapitole 2.4.

Vraťme se však k otázce vyváženosti diachronního korpusu, konkrétně zda by se měly proporce zastoupení jednotlivých variet v čase měnit. Tutéž otázku si klade i Baker, jednoznačnou odpověď ale nedává: „Should corpus builders attempt to capture what people are doing with written language at any given point, or should they create a single

sampling frame and then stick to it? Both options are likely to impact on the sorts of research findings that occur, with the former perhaps resulting in greater differences than the latter.“ (Baker, 2009a, str. 335)

Velice podobná úvaha Oakesova naznačuje, že jde o otázku sice důležitou, ale stále otevřenou: „Another issue arises regarding whether a diachronic corpus be exactly balanced text-for-text, or whether the diachronic corpus should take into account actual cultural differences. In using the same sampling model for the Brown family of corpora, the corpus builders chose the first option. However, it could be argued that perhaps the sampling model should have been changed over time in order to reflect changing patterns of text production and reception. For example, the 1960s could be seen as the heyday of science fiction, with more people writing and buying science fiction than in later decades. Perhaps the composition of a diachronic corpus should take this into account, although this would strongly impact on the amount and type of variation within the corpus.“ (Oakes, 2009, str. 182)

Neměnné složení korpusů celé diachronní řady má řadu praktických výhod, na druhou stranu je tady problém s mizejícími a nově se objevujícími žánry, a také samozřejmě s přesnou vyvážeností takového korpusu, zvlášť pokud by byl časový rozdíl mezi jednotlivými korpusy příliš veliký. Proporce zastoupení jednotlivých variet proměnné v závislosti na čase jsou sice řešením metodologicky čistším, na druhou stranu jsou ale spojené s praktickým problémem, jak stanovit kritéria pro aktualizaci složení korpusu. Není-li navíc časový rozdíl mezi jednotlivými korpusy příliš velký, mohou být malé posuny ve vyváženosti jednotlivých kategorií spíše na obtíž, zvláště z hlediska běžného koncového uživatele, který většinou nemá možnost získané frekvence normalizovat nebo jinak upravovat.

Vezmeme-li jako příklad reprezentativní korpusy řady SYN a jejich složení podle hlavních typů textu, na první pohled nás zarazí výrazný rozdíl ve složení SYN2000 na jedné straně a SYN2005 spolu se SYN2010 na straně druhé. I pokud bychom bez výhrad přijali argumentaci v Králík (2004) a Králík a Šulc (2005), která rozdíl mezi SYN2000 a SYN2005 vysvětluje, je zarážející, že by se potom už vůbec žádná další změna neudála, protože korpus SYN2010 byl záměrně koncipován tak, aby jeho složení ve všech kategoriích *txtype* a *genre* plně odpovídalo složení korpusu SYN2005. Jde tedy o nekonzistentní přístup, který kombinuje obě možnosti popsané v předchozím odstavci. To není šťastná volba zvláště v době, kdy zažíváme boom nových internetových žánrů, se kterými výzkumy konané okolo roku 2000 nepočítaly a ani počítat nemohly. Korpusy řady SYN tak záměrně nezahrnují jazyk primárně internetový, naopak se stále zaměřují pouze na jazyk tištěný, případně jazyk textů sice na internetu publikovaných, ale určených pro tištěná média.

Podobný postoj k internetovým textům zaujímá i Davies (2009), když popisuje složení monitorovacího korpusu COCA zachycujícího americkou angličtinu od roku

1990 do současnosti. V návrhu korpusu byl maximální důraz kladen na jednoduchost a přehlednost, a proto je v něm každý rok zastoupen stejným počtem slov a v jejich rámci je každý z pěti hlavních žánrů (Daviesova terminologie) zastoupen právě jednou pětinou. Davies však nezařazování internetových textů vysvětluje tím, že by se mu je nepodařilo získat v dostatečném rozsahu už od r. 1990 a že v této době některé z nich (např. blogy) ještě vůbec neexistovaly. Důvodem tedy není setrvávání na starším pojetí reprezentativnosti a vyváženosti, těmi se Davies záměrně nezabývá; prioritou je pro něj zachování jednotné architektury korpusu. Je ovšem zřejmé, že přibývání a ubývání jednotlivých hlavních žánrů je problém, se kterým se každý podobný korpus musí dříve nebo později vypořádat. Stejně tak je ale zřejmé, že provázanost jazykových změn se změnami společenskými je dalším argumentem pro nutnost používat jako základ pro diachronní srovnání externí kritéria ve smyslu popsaném v podkapitole 2.4.

3.3 Diachronní srovnávací studie

3.3.1 Druhy diachronních srovnávacích studií

Tato podkapitola se podrobně zabývá studiemi, jejichž cílem bylo popsat s pomocí korpusových dat změny v jazyce a jeho vývoj. Důraz je kladen na použité korpusy, metody a výsledky, a to zvláště ve vztahu k této práci. Vyplyne z nich totiž celkový rámec, do něhož je zasazena, stejně tak jako studie, s nimiž je možné ji srovnávat a v jakých ohledech. Studie jsou rozčleněny do čtyř částí podle druhu korpusů, na nichž jsou založeny, a to na diachronní, synchronní, monitorovací a brownovské.

Pod pojmem ***diachronní korpus*** rozumíme v této práci korpus zachycující jazyk, který již není pociťován jako současný. Jde o korpus vzniklý zpravidla s velkým časovým odstupem od doby vzniku textů, které obsahuje. Protože je sestavován většinou zpětně a obtížně (skenování, korektury, přepisování), je typicky poměrně malý. Naproti tomu ***synchronní korpus*** je odrazem jazyka v době vzniku korpusu, a může proto být (jde-li o korpus psaný) vzhledem k vysokému stupni digitalizace až o několik řádů větší. Termín ***monitorovací korpus*** používáme v dalším textu tak, jak ho chápe Davies (2011), tj. pro korpus většinou synchronní, budovaný přímo s cílem systematického a podrobného zachycování změn jazyka v čase.[1] Takových korpusů je velice málo, protože většina synchronních korpusů je koncipována jako jednorázové. Zvlášť byly vyčleněny ***brownovské korpusy*** vzhledem ke svému výjimečnému postavení, množství studií na nich založených i tomu, že nejde o monitorovací korpusy v pravém slova smyslu.

Pro úplnost dodáváme ještě další možnost, kterou je použití webu nebo webových korpusů pro diachronní srovnání. Ve své případové studii se tímto tématem zabývá Mair

[1] Někteří autoři považují takto koncipovaný korpus za diachronní: „A diachronic corpus is, first and foremost, a collection of texts that vary along the parameter of time.“(Hilpert a Gries, 2009, str. 386)

(2011), jehož postoj k této možnosti je realistický a zdrženlivý. Ačkoli dochází k závěru, že za předpokladu zvýšené opatrnosti lze i pro tyto účely webová data používat, za nejvhodnější považuje kombinovaný přístup propojující web a webové korpusy s korpusy tradičními. Podobně střízlivý postoj zaujímají v obecnějším příspěvku Lüdeling et al. (2007), kteří za nejvýhodnější kompromis považují korpus webový (web-crawled). Poukazují také na fakt, že výhoda velkého množství textů je často vykoupena malou spolehlivostí metadat (rok vydání, typ textu atd.): „While the web is an inherently diachronic resource, it has only existed for a short time span so far, and the available date information is highly unreliable.“ (Lüdeling et al., 2007, str. 15)

V poslední době se navíc s nástupem obrovského množství dat zpřístupňovaných společností Google (Google n-grams, Google Books) objevují také na nich založené popularizační diachronní studie, jakou je například Michel et al. (2011). Projevuje se v nich síla velkých dat, z nichž lze prostřednictvím jazyka snadno získat informace o kultuře a společnosti, které jsou navíc na první pohled zajímavé a přesvědčivé i pro laika. Ukazují také, že množství dat může při pohledu dostatečně daleko do minulosti zakrýt nespolehlivá metadata; za těchto podmínek totiž velké množství zajímavých rozdílů mezi jednotlivými jazykovými jevy vyvstává téměř automaticky, bez ohledu na přesné složení dat nebo spolehlivost jejich zpracování. Tím se však podobné studie vzdalují hlavnímu cíli této práce, jímž je zjistit možnosti a meze na korpusu založeného diachronního srovnání blízkých stavů jazyka. Protože používání velkého množství hrubých internetových dat přináší problémy úplně jiného druhu, budeme se v dalším textu zabývat pouze studiemi založenými na tradičních jazykových korpusech.

3.3.2 Studie založené na diachronních korpusech

Mezi nejznámější diachronní (historické) korpusy angličtiny patří The Helsinki Corpus of English Texts (Kytö, 1991; Rissanen, 1994) a ARCHER (Biber et al., 1994). Oba tyto korpusy byly vytvořeny s cílem umožnit výzkum dlouhodobých vývojových tendencí v angličtině, první z nich pokrývá reprezentativními vzorky období od 8. století do konce 17. století (celkem cca 1,6 mil. slov), druhý od roku 1650 do roku 1990 (celkem cca 1,7 mil. slov), takže se oba korpusy víceméně doplňují. ARCHER však má pravidelnější strukturu, je rozdělen do sedmi padesátiletých období, v každém z nich na deset „žánrů“ po (nejméně) deseti textech, a navíc zahrnuje i vzorky americké angličtiny.

Biber a Finegan (2001) popisují historické změny v angličtině na základě korpusu ARCHER, konkrétně vývoj některých biberovských dimenzí (viz podkapitola 2.3) ve sledovaných psaných registrech od roku 1650 do současnosti. Zjišťují přitom rozdílný vývoj tzv. speech-based registers (dopisy, deníky, beletrie, novinové reportáže), považovaných za aproximaci mluveného jazyka, a registrů odborných (medicína, právo a vědecké texty vůbec): zatímco odborné texty se od mluveného jazyka dále vzdalují,

nespecializované texty určené obecnému publiku se mu naopak začínají přibližovat. Protože podobný vývoj zaznamenal Biber (1995, str. 309–311) i pro somálštinu (kde je navíc z jazykového hlediska velmi specifická situace), nabízí se hypotéza, že by mohl být jazykově nezávislý. Na základě toho proto Biber a Finegan (2001) popisují hypotetické tři hlavní fáze vývoje psaného jazyka: v první fázi, již na počátku písemnictví, je psaný jazyk jasně odlišený od mluveného. Ve druhé fázi se psaný jazyk kultivuje, a tím se dále odlišuje od jazyka mluveného. Konečně ve třetí fázi se některé registry psaného jazyka (odborné texty) jazyku mluvenému dále vzdalují (specializace), texty určené širšímu publiku se mu však začínají naopak přibližovat (demokratizace). Zmíněná demokratizace a přibližování některých registrů psaného jazyka jazyku mluvenému ve třetí fázi přitom odpovídají závěrům řady dalších studií, kterými se budeme zabývat dále.

Jiným typem diachronního srovnání je studie založená na lexikální databázi CELEX (Lieberman et al., 2007) kvantifikující vztah mezi frekvencí anglických nepravidelných sloves a jejich tendencí přecházet k pravidelnému tvoření minulého času a příčestí pomocí sufixu *-ed*. Statisticky tak potvrzuje očekávanou souvislost, že čím je sloveso méně frekventované, tím je jeho přechod k pravidelné konjugaci pravděpodobnější.

Pro češtinu existují studie sledující na základě korpusu DIAKORP vývoj některých pravopisných, fonologických, morfologických, syntaktických a lexikálních jevů, zejména konkurenci variant (Kučera, 2005). Tyto studie inspirovaly vznik obecnějšího nástroje SyD (Cvrček a Vondřička, 2011) umožňujícího studium konkurence libovolných uživatelem zadaných variant, a to nejenom diachronně v závislosti na roku vzniku textu, ale i synchronně. Variabilitu v synchronním smyslu lze studovat v závislosti na typu textu a žánru pro psané texty (korpus SYN2010) a v závislosti na vybraných socioling-vistických parametrech (věk, pohlaví, vzdělání, nářeční oblast) pro mluvené projevy (korpusy ORAL2006 a ORAL2008).

SyD umožňuje plně využívat dotazovací jazyk CQL korpusového manažeru Manatee (Rychlý, 2000), je částečně parametrizovatelný a využívá normalizace zobrazovaných frekvenčních údajů. Jeho diachronní část je založena na korpusu DIAKON (veřejně nepřístupný, v současné době má zhruba 120 milionů slov), který vznikl spojením souboru dosud nezkorigovaných textů diachronní složky ČNK (35 mil. slov) s vybranými publicistickými texty korpusu SYN (80 mil. slov), malou částí beletrie z 80. let z téhož korpusu (1 mil. slov) a Rudým právem z roku 1952 z dosud veřejně nepřístupného korpusu TOTALITA (4 mil. slov). Toto jeho složení je do značné míry dáno především praktickou dostupností textů z jednotlivých časových období. Ačkoli je tedy samotný korpus DIAKON nereprezentativní a spíše panchronní, nad ním stojící nástroj SyD nabízí řadu velice užitečných funkcí umožňujících studium variability jazyka v mnoha směrech.

Na závěr bychom chtěli zdůraznit, že zmíněné diachronní korpusy a na nich založené studie se od našeho přístupu liší ve dvou zásadních věcech: jsou relativně malé, přitom

však pokrývají dlouhé období vývoje jazyka. Naopak řadu SYN tvoří relativně velké korpusy zachycující stavy jazyka velmi blízkých časových období. Tyto rozdíly mají závažné důsledky jak pro volbu srovnávacích metod, tak i pro jejich výsledky: velký korpus sice minimalizuje vliv výběru konkrétních textů, na druhou stranu jsou ale vývojové tendence dokumentované v diachronních korpusech díky velkému časovému rozpětí zřetelné a snadno identifikovatelné. Tento fakt vysvětluje odlišnost přístupu zvoleného v kapitole 5 stejně jako skutečnost, že výsledky poukazují pouze na tendence, nikoli dokončené jazykové změny.

3.3.3 Studie založené na synchronních korpusech

Na rozdíl od korpusů diachronních se při vytváření synchronních korpusů zpravidla nebere zřetel na možnost jejich použití pro zachycování jazykového vývoje ani na srovnatelnost daného synchronního korpusu s jinými: „Many traditional corpora, such as the BNC, are designed to be synchronic, so that diachronic analysis is only possible when a comparable corpus with material from a different time is available.“ (Lüdeling et al., 2007, str. 15). Ačkoli je tedy synchronních korpusů většina, chybějící srovnatelnost alespoň dvou z nich je hlavním důvodem velmi malého počtu na nich založených studií a také – jak vzápětí ukážeme – různorodosti těchto studií. V této souvislosti vyniká unikátnost české situace, kde jsou k dispozici hned tři reprezentativní synchronní korpusy pokrývající tři po sobě jdoucí časová období (SYN2000, SYN2005 a SYN2010). Jejich diachronní srovnání je hlavním tématem této práce i proto, že tyto korpusy prozatím pro podobné účely systematicky využity nebyly.

Asmussen (2006) ve své studii srovnává dva korpusy dánštiny různého složení, Korpus 90 a Korpus 2000. Oba korpusy mají 28 milionů slov, Korpus 90 obsahuje texty z let 1983–1992, zatímco Korpus 2000 z let 1998–2002. Problémem je však rozdílné složení obou korpusů: zatímco dvě třetiny Korpusu 2000 tvoří novinové texty, je jejich podíl na Korpusu 90 pouze třetinový. Tento rozdíl se autor nesnaží kompenzovat normalizací, srovnává prosté frekvence jednotlivých předem vybraných jevů na několika jazykových úrovních (hlavně lexikon, morfologie a kolokace, ale i syntax a sémantika) mezi celými korpusy a dochází k závěru, že vliv rozdílného složení obou korpusů na výsledky může být značný. Jím navrhovaný nástin metodologie pro vyhodnocování diachronní srovnatelnosti korpusů je popisován dále v podkapitole 3.4.

Altintas et al. (2007) vycházeli z korpusu dvojic překladů několika děl z různých jazyků (angličtiny, francouzštiny a ruštiny) do turečtiny. Překlady se od sebe lišily dobou vzniku, rozdíl činil zhruba 50 let. Srovnáváním novějšího a staršího překladu téhož textu zjistili, že se výsledná turecká slova prodlužují a kořeny zkracují. Znamená to, že v turečtině je zřejmá tendence používat ve srovnání s dobou před 50 lety více sufixů, novější překlady navíc používají statisticky významně méně různých kořenů.

Naproti tomu Belica (1996) používá malý, specializovaný korpus o cca 4 milionech slov, složený z nejrůznějších politických textů z let 1989–1990 a rozdělený na západoněmeckou a východoněmeckou část. Začíná vygenerováním všech 1-gramů a 2-gramů, následuje vyhodnocení jejich distribuce po sedmi časových úsecích a výběr těch s největšími rozdíly podle čistého χ^2, χ^2 s Yatesovou korekcí a log-likelihood (významnost každé míry z této trojice je „vážena“ snadno měnitelnými heuristickými koeficienty). V dalším kroku se k nim pomocí MI-score a T-score najdou kolokace, vyhodnotí se jejich distribuce po časových úsecích, v případě významných rozdílů v rozložení se původní n-gramy spolu s kolokacemi přidají do původní množiny n-gramů a celý postup se opakuje. Při vyhodnocení této metody se důraz spíše než na interpretaci výsledků klade na její popis, konfigurovatelnost systému a možnosti jeho různého nastavení, kterými lze výsledky ovlivňovat.

Partington (2010) a Duguid (2010) založili své studie na dvou publicistických korpusech z let 1993 a 2005 nazvaných SiBol 93 (100 mil. slov) a SiBol 05 (144 mil. slov). Korpusy obsahují produkci hlavních britských novinových titulů (The Times, Telegraph a Guardian) a jsou záměrně sestaveny tak, aby byly v maximální míře srovnatelné jako zdroj dat pro MD-CADS (Modern Diachronic Corpus-Assisted Discourse Studies). MD-CADS nabízí jiný úhel pohledu, primárním zájmem je analýza diskurzu prostřednictvím jazyka, v němž se odrážejí společenské a kulturní změny. Metodologicky může jít jak o přístup založený na ověřování předem daných hypotéz, tak i o corpus-driven přístup založený na kvalitativní analýze a kategorizaci klíčových slov, jejichž frekvence v obou korpusech se statisticky významně liší (obdoba metody popisované v kapitole 5). Výsledky ukazují na zvyšující se neformálnost mediálního diskurzu, jeho přibližování bulvárnímu tisku a běžné konverzaci, které se projevují (typicky hyperbolickou) evaluativností, vágností a úbytkem zpráv, jichž je vzhledem k celkovému rozsahu novinových textů stále méně. Tyto závěry se do značné míry shodují se závěry lingvistických studií založených brownovských korpusech, které jsou popsány dále.

3.3.4 Studie založené na brownovských korpusech

Brown University Standard Corpus of Present-Day American English, běžně nazývaný korpus Brown, vznikl začátkem 60. let 20. století na Brown University ve Spojených státech pod vedením W. N. Francise a H. Kučery a obsahuje 1 milion slov psané americké angličtiny z roku 1961. Jeho britským ekvivalentem je korpus LOB (Lancaster–Oslo/Bergen) vytvořený v 70. letech pod vedením G. Leeche, S. Johanssona a K. Hoflanda a obsahující 1 milion slov psané britské angličtiny z roku 1961. Oba korpusy se snaží být reprezentativní a vyvážené, každý z nich se skládá z 500 vzorků vybraných z 15 hlavních textových kategorií, jejichž poměrné zastoupení v každém z korpusů je (až na pár výjimek) shodné. Velikost každého vzorku je okolo 2 000 slov. Také způsob

výběru vzorků byl shodný, šlo o (semi) náhodný výběr textů z kompendií všech textů dané kategorie publikovaných v roce 1961 a o náhodný výběr jednoho vzorku z každého takto vybraného textu. Složení obou korpusů je tedy záměrně co nejpodobnější s cílem umožnit srovnání stavů britské a americké angličtiny v roce 1961 (Hofland a Johansson, 1982).

Tato dvojice korpusů byla počátkem 90. let doplněna korpusy Frown (Freiburg-Brown; americká angličtina z roku 1992) a FLOB (Freiburg-LOB; britská angličtina z roku 1991) vytvořenými podle stejných kritérií pod vedením C. Maira na univerzitě ve Freiburgu. Vznikla tak čtveřice korpusů stejného rozsahu a složení umožňující nejenom srovnání britské a americké angličtiny ve dvou bodech časové osy, ale také diachronní srovnání dvou stavů jazyka s časovým rozdílem třiceti let. Tyto studie a jejich hlavní závěry budou popsány dále. Čtveřici korpusů Brown, LOB, Frown a FLOB budeme dále pro jednoduchost nazývat ***brownovské korpusy***. Vzhledem k postupnosti jejich vzniku a třicetileté mezeře mezi jednotlivými dvojicemi však nejde o korpusy monitorovací v pravém slova smyslu, za vhodnější považujeme dívat se na ně jako na řadu korpusů synchronních reprezentujících stav dané variety angličtiny v jednom konkrétním roce.

Nejpodstatnější nevýhodou brownovských korpusů je jejich velikost; přes shodnou koncepci, složení a veškerou péči věnovanou výběru vzorku je 1 milion slov málo, zvláště pro studie zabývající se slovní zásobou. Jejich užitečnost i dnes, v době webu a korpusů o několik řádů větších, však obhajuje Baker (2009a, str. 335): „A million words may now seem like a tiny amount of data with the current advances in corpus collection, although there is still value in using a carefully balanced and sampled corpus model such as that used in the Brown family in order to investigate linguistic change and variation.“ Také Hundt a Leech (2011) uvádějí jako hlavní výhody brownovských korpusů:

- pečlivě dodržované složení;
- spolehlivost lemmatizace a morfologického značkování, které lze ručně revidovat, což se také u těchto korpusů postupně děje;
- možnost vyčerpávající kvalitativní analýzy všech výskytů zkoumaného jevu;
- možnost získat celý korpus v textové podobě, což u velkých korpusů nebývá z autorskoprávních důvodů možné.

Brownovské korpusy jsou pro různé druhy srovnání velice populární, vznikají také další doplnění této řady. Jde zejména o korpus B-LOB (before-LOB; britská angličtina kolem r. 1931) a připravovaný B-Brown (before-Brown; americká angličtina kolem r. 1931) sestavované s cílem přímé srovnatelnosti záměrně stejným způsobem jako jejich předchůdci. Ve stadiu příprav je podle práce Hundt a Leech (2011) dokonce i stejně koncipovaný korpus britské angličtiny z doby okolo roku 1901, nedávno vznikl

i novější korpus BE06 s texty z let okolo roku 2006 (Baker, 2009a). Tento korpus sice dodržuje stejné složení jako ostatní brownovské korpusy, vznikl však stažením všech textů z internetu. Ačkoli jednou z podmínek pro zařazení textu do korpusu byla existence jeho tištěné podoby, odlišný způsob výběru textů z jejich elektronických verzí má pravděpodobně vliv i na nižší srovnatelnost korpusu BE06 s ostatními korpusy této řady (Baker, 2009b).

Diachronním i synchronním srovnáváním brownovských korpusů se zabývala řada studií zaměřených na různé jazykové jevy. Téměř všechny se však shodují v tom, že zjištěné rozdíly mezi jazykem z let 1961 a 1991 jsou v rozhodující míře způsobené rozvolňováním norem psaného jazyka a jeho přibližováním jazyku mluvenému. Tento závěr ze srovnání brownovských korpusů pregnantně shrnuje Mair: „It seems that a project originally designed to document ongoing grammatical changes in present-day English has in fact produced a record of a sociolinguistic or discourse-historical phenomenon – the 'colloquialisation' of the norms of written English which has taken place over the past thirty years or so. The increasing frequency of the progressive and the *going-to* future, writers' growing willingness to use contracted forms, and to some extent also the shifts noted in the choice between inflectional and periphrastic genitives are not due to the fact that the grammar of the language itself has changed. Rather, these developments show that informal options which have been available for a long time are chosen more frequently today than would have been the case thirty years ago.“ (Mair, 1997, str. 203) A dále: „Very few genuine instances of grammatical change were noted. Most changes observed could be interpreted as a result of the colloquialisation of the norms of written English which has taken place over the past thirty years. This colloquialisation is the linguistic correlate of a general social trend towards greater informality.“ (Mair, 1997, str. 206)

Podrobnějším průzkumem rysů způsobujících větší neformálnost jazyka se dále zabývají Hundt a Mair (1999), o „colloquialisation and democratisation“ píše také Baker (2009a, str. 330). Podobně charakterizuje možné příčiny pozorovaného výrazného poklesu frekvence téměř všech anglických modálních sloves také Leech (2003, str. 236–237) jako „colloquialization“, „democratization“ (konkrétně posun v tom, které modální významy lidé vyjadřují) a „americanization“ (přizpůsobování se trendům pocházejícím z americké angličtiny); ke stejným závěrům dochází i v další studii (Leech, 2004). Na tomto místě je ovšem třeba poznamenat, že s některými dílčími Leechovými závěry, konkrétně s frekvenčním poklesem téměř u všech modálních sloves, nesouhlasí Millar (2009), jeho výsledky na TIME korpusu (viz oddíl 3.3.5) jim totiž neodpovídají. I tato data však potvrzují obecný trend směrem ke vzrůstající neformálnosti psaných textů.

3.3.5 Studie založené na monitorovacích korpusech

Dostáváme se ke složení a způsobům využívání korpusů koncipovaných jako monitorovací, tj. s cílem systematického zachycování změn jazyka v čase; konkrétně jde o korpusy COCA, COHA a TIME sestavené pod vedením Marka Daviese na Brigham Young University ve Spojených státech. Tyto korpusy jsou (spolu s dalšími, nemonitorovacími) přístupné ze stránky `http://corpus.byu.edu/`. Podstatným rysem celého systému je jednotné uživatelské rozhraní navržené s cílem usnadnit výzkum variability jazyka ze synchronního a diachronního pohledu. Podporovány jsou tedy například dotazy srovnávající kolokace daného výrazu v různých hlavních varietách (podle Daviesovy terminologie „genres“) nebo časových obdobích, což je v řadě jiných systémů možné jen s velkými obtížemi (pokud vůbec). Všechny frekvence jsou přitom udávány i normalizované na milion slov, a tedy vzájemně srovnatelné. Netradiční je také serverová část založená na relačních databázích (Davies, 2009), zatímco většina evropských systémů používá specializované korpusové manažery, jakými jsou například Manatee, CQP, Xaira nebo Poliqarp.

Korpus COCA (Corpus of Contemporary American English) je pravidelně doplňovaný monitorovací korpus obsahující americkou angličtinu od roku 1990 do současnosti. Při popisu jeho složení vyzdvihuje Davies (2009) zejména jednotnou architekturu se stejným zastoupením hlavních žánrů v každém roce: každý rok je v korpusu zastoupen přibližně 20 miliony slov, a v jejich rámci je každý z pěti hlavních žánrů zastoupen 4 miliony slov. Těmito hlavními žánry jsou spoken, fiction, popular magazines, newspapers a academic journals, každý rok se přitom doplňují z víceméně stejných zdrojů. Velkou výhodou tohoto přístupu je srovnatelnost dat, ačkoli jsou jednotlivé žánry dále složené z tzv. podžánrů, jejichž podíly jsou velice vágně charakterizovány jako „good mix“ (Davies, 2009, str. 161–162). Autor přitom neuvádí, co to přesně znamená, ani zda považuje zastoupení jednotlivých podžánrů za vyvážené. Reprezentativnost korpusu COCA jako celku je sporná i proto, že způsob výběru a získávání dat se zdají být oportunistické v tom smyslu, že do korpusu jsou zařazeny nejsnáze dostupné texty, které se zároveň hodí do dané kategorie. To je vidět hlavně na mluvené složce, která obsahuje výhradně přepisy televizních a rozhlasových pořadů (i když Davies zdůrazňuje jejich improvizovaný a konverzační charakter), ale i na beletrii, která je z velké části zastoupena časopisecky vydanými texty a filmovými scénáři (Davies, 2009, str. 161).

Podobné složení jako COCA má také korpus COHA (Corpus of Historical American English) obsahující více než 400 milionů slov z let 1810–2009 (Davies, 2011). Každé desetiletí je v korpusu zastoupeno celkem čtyřmi žánry v zásadě odpovídajícími těm v COCA, ovšem bez mluveného jazyka. Proporce jednotlivých žánrů se sice příliš nemění v čase, nejsou však stejné: zhruba polovinu tvoří fiction, zatímco popular magazines, newspapers a non-fiction books se dělí o tu druhou. Také množství dat z jed-

notlivých desetiletí je výrazně rozdílné, celkový objem textů z let 1810–1819 je oproti letům 2000–2009 zhruba 4%. Reprezentativnost korpusu COCA je tedy také snadno zpochybnitelná, což je ovšem u diachronních korpusů běžný problém daný především obtížnou dostupností vhodných textů.

Přes výše uvedené výhrady je dvojice korpusů COCA a COHA ve světovém měřítku jedinečná, zvláště ve spojení s vyhledávacím rozhraním podporujícím studium jazykových změn a umožňujícím zobrazování frekvenčního průběhu uživatelem zadaných slov, slovních spojení nebo kolokací v čase a po jednotlivých žánrech. Možnosti rozhraní ukazuje Davies (2010) na korpusu COCA na několika jazykových úrovních.

V této souvislosti bychom chtěli znovu poukázat na problém reprezentativnosti monitorovacích korpusů, zejména na vztah mezi jazykovými a společenskými změnami a jejich projevy ve složení korpusu. Jak již bylo zmíněno v podkapitole 3.2, je už na teoretické úrovni velice těžké je od sebe oddělovat. Příklad složení korpusů COCA a COHA však ukázal, že při sestavování korpusů často převáží praktické problémy jiného druhu. Zvolená řešení proto musejí být pragmatická, využívající například snadné dostupnosti textů určitého typu, přestože jejich vliv na výstupy založené na takto vzniklém korpusu může být značný.

Dalším z řady korpusů zpřístupněných Markem Daviesem je korpus TIME. Obsahuje kompletní vydání týdeníku TIME od roku 1923, kdy začal vycházet, až do současnosti. Jeho celková velikost přesahuje 100 milionů slov, každý rok je zastoupen 1–1,5 miliony slov. Tento korpus používá Millar (2009) pro svou studii zaměřenou primárně na modální slovesa, avšak s metodologickými dopady na diachronní srovnání vůbec. Zdůrazňuje přitom výhodu používání relativně omezených, ale homogenních a jasně definovaných dat: „although the claims that can be made are necessarily limited, they are securely grounded“ (Millar, 2009, str. 216). Uvádí také následující konkrétní výhody při používání korpusu TIME pro podobné studie, zvláště ve srovnání s korpusy brownovskými (Millar, 2009, str. 193):

- velikost;
- homogenita („internal consistency“);
- jasně daná reprezentativnost: 100% vzhledem k TIME, dostatečná vzhledem k „American newsmagazines“, sporná vzhledem k jazyku jako celku; právě v publicistice se však nejčastěji objevují jazykové změny, viz např. Hundt a Mair (1999);
- pokrytí téměř osmdesáti let;
- pouze roční granularita tohoto pokrytí.

Jedním z cílů studie je srovnání tendencí ve vývoji anglických modálních sloves získaných na korpusu TIME se závěry Leechovými (Leech, 2003) založenými na brownovských korpusech. Millar totiž dochází k výsledkům, které jsou v některých ohledech výrazně odlišné, tyto rozdíly zkoumá a dochází k závěru, že tyto odlišnosti nejsou způsobeny rozdíly ve složení korpusů, protože obdobné (i když o něco menší) diskrepance lze najít i při srovnání novinového subkorpusu Brown/Frown s TIME subkorpusem pouze z let 1961 a 1991 (roky pokryté korpusy Brown a Frown, viz oddíl 3.3.4). Důvodem rozdílů je zřejmě relativně velká frekvenční oscilace některých modálních sloves v čase spojená se srovnáváním pouze ve dvou časových okamžicích (což je dáno dostupností dat brownovských korpusů) a s tím, že brownovské korpusy jsou z hlediska dnešních standardů velice malé, a tedy ve značné míře závislé na zařazení konkrétních textů (vzorků; jedním z důvodů výše zmíněné oscilace přitom může být i toto vzorkování). Autor také uvádí konkrétní příklad, kdy výsledky pozorování konaného pouze pro roky 1961 a 1991 odporují výraznému celkovému trendu potvrzenému od r. 1923 do současnosti. Tato zjištění podtrhují potřebu používat pro podobné diachronní studie velké korpusy a více časových okamžiků.

3.4 Vliv složení korpusu na interpretaci výsledků

Jedním z cílů kapitoly 3 bylo poukázat na nedostatek dat vhodných pro diachronní srovnání. Ideálem je mít k dispozici takový monitorovací korpus (případně řadu korpusů reprezentujících po sobě následující stavy jazyka), ve kterém by rozdíly mezi jednotlivými obdobími (korpusy) byly způsobeny pouze změnami v jazyce, a to včetně změn společenských tak, jak je jazyk odráží a jak jsou v něm zachyceny. Tento ideál je ale nedosažitelný: už kvůli problémům s reprezentativností a vyvážeností synchronního korpusu v jednom časovém okamžiku zmiňovaným v kapitole 2 je prakticky nemožné dosáhnout plné reprezentativnosti monitorovacího korpusu ve všech časových obdobích. Zvláště při srovnávání blízkých, a tedy velice podobných stavů jazyka jsme navíc vystaveni riziku zvýšeného vlivu různého složení korpusu na výsledky tohoto srovnání, takže lze jen velmi obtížně posuzovat, do jaké míry zjištěné rozdíly skutečně odpovídají změnám v jazyce. Pro dosažení maximální možné diachronní srovnatelnosti je tedy potřeba minimalizovat vliv výběru textů do korpusu, přičemž ovšem opět nastává metodologický problém analogický interním a externím kritériím pro stanovení reprezentativnosti (viz podkapitola 2.4), a to, zda používat interní nebo externí kritéria pro stanovení míry diachronní srovnatelnosti korpusů.

Asmussen (2006) navrhuje najít invariant, časově nezávislý jazykový jev nebo spíše množinu jevů, jejichž frekvence se mezi dvěma blízkými časovými obdobími nemění nebo se mění jen velice málo. Tento invariant by se pak používal jako korektiv měřící vhodnost korpusů pro diachronní srovnání. Ačkoli v článku o některých možných inva-

riantech diskutuje, dochází k závěru, že jejich volba není jednoduchá už z toho důvodu, že není vůči čemu ověřovat jejich skutečnou invariantnost. Nebezpečí jejich neinvariantnosti je příliš velké, protože snaha udržet tyto neověřené invarianty konstantní by mohla vést k zastření skutečných posunů v jazyce nebo naopak k interpretaci invariantů skutečných jako jazykových změn. Výpovědní hodnota výsledků srovnání korpusů s takto definovanou srovnatelností by se tak relativizovala především k těmto invariantům samým, o jejich možné interpretaci vzhledem ke změnám v jazyce by tedy sama o sobě neříkala nic.

Přes nespornou užitečnost a objektivnost interních kritérií, jakými jsou například výše uvedený invariant nebo Biberova faktorová analýza (viz podkapitola 2.3), používaných jako zpětná vazba budovaných nebo srovnávaných korpusů, je nelze používat jako rozhodující kritérium pro stanovení diachronní srovnatelnosti, protože sama o sobě nemohou odlišit rozdíly ve složení korpusů dané časem a složením korpusu. Zdá se tedy, že se nelze vyhnout intuici jako konečné instanci při posuzování nejenom srovnatelnosti, ale také výsledků diachronního srovnání korpusů. Zvláště při srovnávání blízkých stavů jazyka, které je tématem této práce, je však zároveň největší i riziko, že vliv složení korpusu bude silnější než jakákoli jazyková změna: „It could be the case, however, that some of the results can be better explained by corpus building procedures rather than cultural changes.“ (Baker, 2009a, str. 334), a dále „I suspect that some findings may be linked to certain types of texts that were included in the corpus (autobiographies, academic papers on health or children), although it is difficult to ascertain whether these trends are due to changes in society, or due to their over-inclusion despite my attempts at random sampling.“ (Baker, 2009a, str. 335). Baker tady mluví o malých, pouze milionových brownovských korpusech a korpusu BE06 (viz oddíl 3.3.4), zatímco při používání stomilionových korpusů je toto riziko jistě menší, přesto však reálné. Ačkoliv je samozřejmě možné příčiny pozorovaných změn ve zdrojových korpusech dále ověřovat, při jejich hodnocení se jisté míře subjektivity vyhnout nelze. O nutnosti obezřetné interpretace výsledků získaných srovnáváním korpusů při jejich zobecňování na jazyk píše řada autorů, toto téma bylo již zmíněno také v podkapitole 2.2. Příklady je možné najít v příslušných částech kapitoly 6, v níž prakticky demonstrujeme problémy reprezentativnosti a diachronní srovnatelnosti, které jsme dosud popisovali pouze teoreticky.

4 Popis zdrojových dat

4.1 Korpusy řady SYN

Tato kapitola podrobně popisuje synchronní reprezentativní korpusy řady SYN použité jako zdrojová data pro metodu popisovanou v kapitole 5. Vzhledem k tomu, že cílem práce je zjistit možnosti a meze detekce tendencí jazykového vývoje založené právě na těchto korpusech, považujeme jejich důkladný popis za nezbytný. Jak již bylo uvedeno v podkapitole 3.4, je zvláště při diachronním srovnání korpusů zachycujících jazyk velice blízkých období důležité podrobně znát jejich složení a způsob zpracování, které mohou mít na výsledky podstatný vliv. To bude také potvrzeno později v diskusi o výsledcích těchto metod, nezbytnost podrobného seznámení se zdrojovými daty zdůrazňuje také Partington (2010, str. 90).

Synchronní psané korpusy jsou v ČNK zastoupeny především korpusy řady SYN. Hlavním cílem této řady, jejíž základy byly položeny již při vzniku ÚČNK, je kontinuální mapování stavu a vývoje současné psané češtiny. Řadu SYN tvoří v současné době pětice korpusů SYN2000, SYN2005, SYN2006PUB, SYN2009PUB a SYN2010, vročení každého z nich udává (s výjimkou korpusu SYN2009PUB) rok zveřejnění. Základní charakteristiky těchto korpusů jsou uvedeny v tabulce 4.1.1.

	velikost	základní charakteristika
SYN2000	100 mil.	reprezentativní; převažují texty z let 1990–1999
SYN2005	100 mil.	reprezentativní; převažují texty z let 2000–2004
SYN2006PUB	300 mil.	publicistický; pouze texty z let 1989–2004
SYN2009PUB	700 mil.	publicistický; pouze texty z let 1995–2007
SYN2010	100 mil.	reprezentativní; převažují texty z let 2005–2009

Tabulka 4.1.1: Přehled korpusů řady SYN.

Základ řady SYN tvoří reprezentativní, vyvážené stomilionové korpusy SYN2000, SYN2005 a SYN2010 (podrobnému rozboru jejich reprezentativnosti je věnována podkapitola 4.3) zveřejňované vždy po pěti letech a zachycující psanou češtinu tří po sobě následujících časových období. Za doplněk těchto reprezentativních korpusů lze považovat publicistické korpusy SYN2006PUB a SYN2009PUB složené výhradně z publicistic-

kých textů z let uvedených v tabulce 4.1.1 (texty z roku 1989 tvoří pouze porevoluční Informační servis). Tyto korpusy si nečiní žádné nároky na reprezentativnost, a to ani v rámci publicistiky, a také jejich velikost a rok zveřejnění jsou nepravidelné, neboť jsou dány především množstvím a dostupností textů. Hlavní motivací jejich vzniku bylo poskytnout velké množství dat uživatelům, kterým z různých důvodů nestačí rozsah reprezentativních psaných korpusů.

Konkrétní složení každého z korpusů řady SYN (reprezentativního nebo publicistického) je výsledkem výběru textů z jednotné banky synchronních psaných textů. Tato ***banka*** (Kocek et al., 2000, str. 20) se buduje průběžně a tvoří ji všechny texty, které již prošly všemi fázemi zpracování a jsou připravené k zařazení do některého z korpusů. V této souvislosti poznamenejme, že každý z textů v bance může být zařazen nejvýše do jednoho korpusu tak, aby celá řada SYN obsahovala pouze disjunktní korpusy. Podstatným momentem tu však je, že všechny texty procházejí před zařazením do banky záměrně stejnou sekvencí procedur, které je především normalizují a čistí. To se děje bez ohledu na pozdější zařazení textu do některého z korpusů řady SYN, které ostatně není v okamžiku zpracování textu známo.

4.2 Technické aspekty výstavby korpusů řady SYN

4.2.1 Terminologický úvod

Cílem této podkapitoly je popsat všechny procedury, kterými procházejí texty před zařazením do některého z korpusů řady SYN, a poukázat přitom na některé důsledky tohoto postupu, které jsou podstatné pro rozbor výsledků v kapitole 6. Tento popis považujeme za potřebný i proto, že se text původní korpusové příručky (Kocek et al., 2000) týká pouze korpusu SYN2000, je již celkově zastaralý a žádná jeho dostatečně podrobná aktualizace se od té doby neobjevila. Proto se také v následujícím popisu nutně objevují tvrzení, která se neopírají o literaturu, ale jen o osobní zkušenosti a údaje autora, který se po mnoho let zabýval jak řadou těchto procedur, jejich programováním a dolaďováním, tak i výběrem textů do všech korpusů řady SYN. Celý postup zpracování textů od jejich získání po zveřejnění v některém z korpusů řady SYN je možné rozdělit do několika fází, každé z nich je věnována samostatná část textu.

Než se k nim však dostaneme, je nutná terminologická odbočka. V korpusové lingvistice jsou etablovány anglické pojmy ***type*** (slovní tvar jako takový) a ***token*** (konkrétní výskyt daného slovního tvaru v korpusu). Platí tedy, že frekvence daného typu je dána počtem všech jeho tokenů v korpusu a že celkový součet všech tokenů všech typů dává velikost korpusu. Poznamenejme, že se tato distinkce nemusí týkat jen slovních tvarů, může jít právě tak i o lemmata (základní slovní tvary) nebo morfologické značky.

Kromě těchto obecně známých termínů budeme dále v souvislosti s korpusovým manažerem Manatee/Bonito (Rychlý, 2000; Rychlý, 2007) používat některé termíny speciální. Pro Manatee/Bonito je korpus sledem unikátně očíslovaných ***pozic***, které v zásadě odpovídají tokenům a které jsou základními jednotkami pro vyhledávání v korpusu (Rychlý, 2000, str. 9). Pozici typicky tvoří slovní tvar, číslo nebo interpunkční znaménko, které bylo při tokenizaci (viz dále) osamostatněno. Celkový počet pozic udávaný korpusovým manažerem je tedy vždy vyšší než velikost korpusu v počtu slov udávaná na stránkách ÚČNK nebo v podkapitole 4.1, kde se slovem rozumí pouze pozice obsahující alespoň jeden alfabetický znak.

Pro každý korpus je dále určena sada ***pozičních atributů***, jejichž hodnoty musejí být pro každou pozici definovány. Minimální sadou je jediný poziční atribut nazývaný ***word***, jehož hodnotou je zpravidla slovní tvar tak, jak se vyskytl v textu. Je však samozřejmě možné ke každé pozici doplnit další poziční atributy, například pro všechny korpusy řady SYN jsou kromě implicitního pozičního atributu *word* definovány ještě poziční atributy ***lemma*** a ***tag***. Ty tvoří doplňkovou informaci, kterou není potřeba využívat a jejíž přidání nijak neovlivňuje původní text. Jednotlivé pozice od sebe mohou být odděleny ***strukturními značkami***, které korpus hierarchicky rozčleňují. V případě korpusu SYN2010 jde o strukturní značky *opus*, *doc* a *s* odpovídající členění na jednotlivá díla (*opus*), která se skládají z dokumentů (*doc*; typicky jde o články nebo kapitoly) a ty zase z vět (*s*); každá pozice přitom „patří“ do konkrétní věty, dokumentu a díla. Tyto strukturní značky jsou přitom často dále opatřeny tzv. ***strukturními atributy***, například strukturní značka *opus* má v korpusu SYN2010 mimo jiné definovány strukturní atributy *autor* (autor díla), *nazev* (jeho název), *rokvyd* (rok vydání) nebo *id* (jednoznačný identifikátor díla).

4.2.2 Akvizice textů

Před zahájením zpracování je třeba kromě vlastního textu získat i souhlas s jeho použitím v ÚČNK, který je stvrzen podepsáním dohody o spolupráci. Přitom je nezbytné aktivní oslovování nakladatelů, aby dodávali další texty podle již uzavřených smluv, případně uzavírání smluv nových; to se však nemusí vždy podařit, někteří nakladatelé texty poskytnout odmítají. Po podepsání dohody následuje předání elektronické podoby textů do ÚČNK. Další možností je stahování textů z internetu, které se však používá spíše doplňkově. Hlavním důvodem je fakt, že po získání souhlasu s použitím textů nebývá problém texty získat jednodušeji a v lepší kvalitě přímo od vydavatele, protože čištění HTML kódu není příliš spolehlivé. Tento postup také většinou umožňuje snadnější zpracování i co se týče přesné identifikace konkrétních čísel jednotlivých periodik. Jiné způsoby získávání textů, které zahrnují nutnost převodu do elektronické podoby (skenování nebo manuální přepisování), se používají již jen výjimečně.

Tady bychom chtěli upozornit na fakt, že ačkoli je výběr textů do reprezentativních korpusů záměrný (viz podkapitola 4.3), jde o výběr závislý na možnostech banky, a tedy na tom, jaké texty se do ní podaří získat. Existují totiž druhy textů lépe a hůře dostupné, jednoznačně nejsnazší je (vzhledem k výslednému objemu dat) získat periodika, protože jde většinou o velký balík textů v jednotném formátu. To je také hlavní důvod, proč řada světových korpusů obsahuje převážně periodika, a z důvodu přebytku periodik (zejména novin) vznikly i korpusy SYN2006PUB a SYN2009PUB.

4.2.3 Konverze do meziformátu

Protože jsou takto získané texty v nejrůznějších formátech (prosté textové soubory, wordovské dokumenty, PDF, výstupy DTP programů jako QuarkXPress aj.), je nezbytným dalším krokem jejich sjednocení. K tomu dojde konverzí do jednotného meziformátu, kterým je prostý ASCII text v kódování CP1250 (používání osmibitového kódování češtiny je dáno historicky), v němž jsou použity SGML entity pro zápis znaků z jiných znakových sad (např. *è* namísto *è*) a obecně jevů, které nelze v prostém textu zapsat, ale které je vhodné zachovat (např. tučné písmo, kurzíva, horní a dolní index apod.). V této části konverze také musejí být objekty netextového charakteru (obrázky, grafy apod.) odstraněny, a případně také nahrazeny entitami označujícími jejich vypuštění. Totéž platí i o částech charakteru sice textového, na které se však z hlavního textu pouze odkazuje, takže by buď musely zůstat mimo něj, nebo by jejich zařazení plynulost textu narušilo (krátké poznámky pod čarou, popisky k obrázkům apod.).

Konverze do meziformátu se provádějí v zásadě automaticky, vlastním konverzím ovšem zpravidla předcházejí ještě adaptace a parametrizace konverzních programů, které jsou pro jednotné zpracování jevů specifických pro konkrétní balík textů výhodné právě v této fázi.

4.2.4 Anotace

Po konverzi do meziformátu probíhá ***anotace*** všech textů (Kocek et al., 2000, str. 25), která v sobě zahrnuje jak anotaci bibliografickou, tak i evaluativní určování typu textu a žánru, na které se v dalším textu soustředíme. Anotace se provádí buď jednotlivě (neperiodika), nebo dávkově (periodika; ty však v bance převažují, jejich podíl na celkovém množství textů tvoří zhruba 80 %). Protože je anotace jednotlivých neperiodik práce převážně manuální, nutně spojená se zevrubným prohlížením anotovaných textů, je její nedílnou součástí také další čištění textu, vypuštění obsahu, tiráže, rejstříku a jiných seznamů, obecně takových částí textu, které jsou heslovité a neposkytují dostatečný kontext.

Typ textu a ***žánr*** jsou hodnoty klíčové pro pozdější vyvažování reprezentativních korpusů (viz oddíl 4.2.6), a proto je také složení textů v bance s ohledem na tyto hodnoty pravidelně vyhodnocováno a slouží jako zpětná vazba pro akvizice textů. Zároveň jsou obě hodnoty zveřejněny jako strukturní atributy ***txtype*** a ***genre***, které lze v korpusovém manažeru Manatee/Bonito zobrazovat nebo podle nich vyhledávat. Kromě nich přibyl počínaje korpusem SYN2010 nový atribut ***txtype_group*** obsahující hlavní typ textu s jednou ze tří hodnot *beletrie*, *odborná* nebo *publicistika*. Ty lze sice jednoznačně odvodit z hodnoty *txtype*, vzájemný vztah mezi *txtype* a *txtype_group* však není zřejmý, a proto ho také uvádíme v tabulce 4.2.1. Tabulka navíc ukazuje vztah mezi *txtype_group* a pojmy ***imaginativní text*** (*beletrie*) a ***informativní text*** (*odborná* nebo *publicistika*), které budeme používat dále.

txtype_group	*txtype*	kategorie
		IMAGINATIVNÍ TEXTY
beletrie		
	VER	básně
	SON	písně
	SCR	dramatické texty, scénáře
	NOV	romány
	COL	povídky
	FAC	literatura faktu
	IMA	jiné imaginativní texty
		INFORMATIVNÍ TEXTY
odborná		
	SCI	vědeckonaučná literatura
	POP	populárněnaučná literatura
	TXB	učebnice
	ENC	abecedně, systematicky a jinak uspořádaná díla
	ADM	administrativa
publicistika		
	PUB	publicistika (noviny a neodborné časopisy)
	MIS	rozmanité (efemera)

Tabulka 4.2.1: Vztah atributu *txtype* k hlavnímu typu textu.

Atribut *genre* je naproti tomu na *txtype* nezávislý, i když jisté závislosti de facto tady jsou. Některé žánry totiž mají smysl jenom v kombinaci s imaginativními, nebo naopak

informativními typy textu, jak je vidět z tabulky 4.3.1 na straně 44. V téže tabulce je možné najít také rozčlenění jednotlivých kategorií textu podle hodnot *txtype* a *genre* včetně jejich procentuálního zastoupení, podle něhož byly vyvažovány reprezentativní korpusy. Tabulka je kompletním výstupem výzkumů o reprezentativnosti popisovaných v podkapitole 4.3 a lze ji chápat také jako podrobné rozčlenění základních kategorií z tabulky 4.2.1.

4.2.5 Dočištění a zařazení do banky

Zařazení textu do banky znamená především jeho převod do formátu SGML (volba SGML formátu je opět dána historicky), ve kterém jsou všechny texty v bance uloženy. Ještě předtím však probíhá několikastupňové plně automatické dočištění, z něhož lze vyčlenit čtyři základní procedury. Tyto procedury se v plném rozsahu provádějí pouze při zpracování textů informativních, tedy takových, kterým byla při anotaci přidělena hodnota *txtype* pro publicistiku nebo odbornou literaturu. Má se totiž za to, že z technického hlediska jednoduchý a přehledný beletristický text je již po čištění v anotační fázi připraven k zařazení do banky, a že by mu tedy plně automatické dočištění mohlo spíše uškodit.

Zmíněné dočištění informativních textů v sobě zahrnuje především 1) detekci a odstraňování cizojazyčných částí textu, které se provádějí na úrovni odstavců (krátké několikaslovné citáty tedy zůstávají zachovány). Dále probíhá 2) detekce a odstraňování duplicitních článků z periodik, avšak pouze lokálně, tj. v rámci jednoho titulu a omezeného počtu po sobě následujících čísel. Jde o proceduru pracující na úrovni celých dokumentů, která při zjištění přílišné podobnosti dvou článků odstraní vždy ten kratší.

Kromě toho se provádí 3) detekce odstavců obsahujících příliš mnoho „podezřelých“ rysů, jakými jsou například velké množství čísel, interpunkce, nečeských znaků nebo naopak málo znaků s diakritikou. Na základě důkladně otestovaných mezních hodnot těchto rysů a jejich kombinací jsou pak odstraněny celé odstavce obsahující netextové části, které jsou v odborné literatuře a publicistice běžné, ale jejich přítomnost v korpusu by byla spíše kontraproduktivní; jde především o odstavce plné čísel, tabulky, seznamy a části textu z různých důvodů defektní. Podle nepublikované interní zprávy ÚČNK ze srpna 2001, která tuto proceduru vyhodnocovala, je tak odstraněno celkem asi 5 % textu, v nichž se ale nachází téměř 13 % interpunkčních znamének a více než 50 % číslic. Uvedené výsledky dokumentují nejenom úspěšnost tohoto čištění, ale také jeho potřebnost. Zdůrazněme, že se jedná opět o čištění na úrovni odstavců, tj. odstavec je v textu buď ponechán beze změny, nebo je celý odstraněn.

Konečně probíhá 4) automatické spojování slov omylem rozdělených pomlčkou nebo spojovníkem. Konkrétně jde o spojování slov v případech, kdy ani jedno z rozdělených slov není ve slovníku, zato v něm však je slovo vzniklé jejich spojením. Původní

rozdělený tvar slova zůstává zaznamenán, takže je vždy možné se k němu případně vrátit. Jde o jediný zásah uvnitř odstavce, a proto považujeme za potřebné zdůraznit, že cílem této procedury je oprava zjevných chyb vzniklých sazbou nebo konverzí, v žádném případě nejde o opravy původního textu. Přestože se chyby tohoto typu týkají v průměru pouze každého 10 000. slova (údaj vychází z nezveřejněného vyhodnocení této procedury, které proběhlo v roce 2001), je jejich koncentrace v textech z některých zdrojů mnohonásobně vyšší a jejich výskyt by byl rušivý nejenom ve výsledných konkordancích, ale i při dalším automatickém zpracování textů, zejména lemmatizaci a morfologickém značkování.

Obecně platí zásada, že texty opravujeme pouze tehdy, je-li takový zásah ospravedlnitelný jako oprava chyby, která nebyla záměrem autora a pravděpodobně jím nebyla ani způsobena. Protože však hranice mezi záměrem autora a chybou vzniklou při technickém zpracování textů nemusí být snadno rozpoznatelná, je spojování slov konzervativní (opatrné) v tom smyslu, že k němu dochází pouze v případech, kdy je autorský záměr velice nepravděpodobný. Pro úplnost ještě dodáváme, že se překlepy neopravují vůbec, i když v první řadě proto, že neexistují dostatečně spolehlivé automatické metody. Ačkoli tedy uznáváme hodnotu původních textů, které by měl korpus v ideálním případě pouze monitorovat, je z popisu čistících procedur zřejmé, že texty se do korpusu nedostávají v nezměněné podobě a že by to ani nebylo žádoucí.

Závěrem dodejme, že procedury 3) a 4) byly zavedeny až v roce 2001, jejich nezbytnost se totiž ukázala až po zveřejnění korpusu SYN2000. To vzhledem k neměnnosti tohoto korpusu znamená, že jimi – na rozdíl od všech ostatních korpusů řady SYN – neprošel žádný z textů v korpusu SYN2000.

4.2.6 Výběr textů z banky

Banka je koncipována jako repozitář bibliograficky anotovaných a dočištěných textů v jednotném SGML formátu určený pro jejich dlouhodobé uložení. Výběr textů do některého z korpusů řady SYN proto probíhá přímo z banky a při další práci už není potřeba se vracet k jejich původní podobě.

Způsob výběru textů do publicistických korpusů SYN2006PUB a SYN2009PUB je dán především možnostmi banky a spočívá víceméně pouze v označení dosud nezveřejněných publicistických textů požadovaného rozsahu. V dalším textu tohoto oddílu se proto budeme zabývat pouze způsobem výběru textů do reprezentativních korpusů SYN2000, SYN2005 a SYN2010, který nazýváme ***vyvažování***. Vyvažování je převážně manuální výběr z dosud nezveřejněných textů v bance do reprezentativního korpusu, jehož cílem je takové složení textů, které by co nejvíce odpovídalo konkrétním proporcím jejich ideálního zastoupení daným výsledky výzkumů o reprezentativnosti; prakticky jde o výběr textů podle kategorií uvedených v tabulce 4.3.1 na straně 44. Koncept

reprezentativnosti, na němž je tabulka založena, popisujeme v podkapitole 4.3; v tomto okamžiku je podstatné hlavně to, že pro každou kategorii je jednoznačně dáno množství slov ve korpusu a že tato kategorie je určená hodnotami dvojice značek *txtype* a *genre*.

Poznamenejme, že z tabulky vyplývající vazba konkrétních kategorií na *txtype* a *genre* nebyla vždy jednoznačně definována pro všechny jejich kombinace. To si později vyžádalo doplnění těchto vazeb k výstupům vzešlým z výše zmíněných výzkumů, po němž mohla teprve vzniknout definitivní podoba tabulky. Vznikla tak však nejednotnost mezi jednotlivými korpusy v tom, do kterého *txtype_group* by se měly počítat kombinace publicistického typu textu a odborného žánru. Například kombinace *txtype=PUB* (publicistika) a *genre=ARC* (architektura) byla při vyvažování korpusu SYN2000 považována za odbornou literaturu (převážilo hledisko žánru), zatímco při vyvažování korpusů SYN2005 a SYN2010 již byla považována za publicistiku (převážil typ textu); tuto nejednotnost vyjadřuje sloupec „striktní *txtype*“ v tabulce 4.3.4 na straně 50.

Cílem vyvažování tedy je každou kategorii danou hodnotami *txtype* a *genre* naplnit co nejpřesněji daným počtem slov. V případě, že požadované množství slov pro určitou kategorii není v bance k dispozici, je jí odpovídající podíl rozpočítán mezi kategorie sousední. K tomu však dochází spíše výjimečně a pouze u kategorií nejnižší úrovně, takže tím reprezentativnost a vyváženost korpusu na vyšších úrovních není dotčena.

Výsledky výzkumů však neříkají nic o tom, jak by měl probíhat výběr textů v rámci těchto kategorií. V praxi proto vznikla sada dodatečných, nezřídka protichůdných kritérií pro výběr textů uvnitř každé kategorie, k nimž se při vyvažování přihlíží:

- čtenost: základní požadavek pro reprezentativnost založenou na recepci, který však často jde proti literární kvalitě;
- stáří textu: preference novějších textů nemusí vždy odpovídat čtenosti;
- původnost: preference původních českých textů opět nemusí být v souladu se čteností;
- pestrost: snaha o zařazení co nejvíce různých autorů, témat, produkce různých nakladatelství atd.;
- rozsah: důležité zvláště v případě malých kategorií, které by neměly být reprezentovány jedním nebo dvěma rozsáhlými texty; více kratších textů prospívá pestrosti;
- technická kvalita: jde hlavně o předpokládané množství chyb v textu dané jeho zdrojem a konverzemi.

Konkrétním výstupem vyvažování je tedy pro každou kategorii množina textů, která – v rámci možností daných bankou – splňuje také tato kritéria. Zdůrazněme proto, že vyvažování je převážně manuální a v této podobě nesnadno automatizovatelné. Výsledky

náhodného výběru textů z banky by přitom nebyly uspokojivé už kvůli tomu, že zastoupení jednotlivých textů v bance není v žádném případě reprezentativní a neodpovídá jejich čtenosti ani jiným kritériím.

Zvláštním případem je vyvažování publicistických textů, zejména proto, že tvoří velkou a nijak dále nečleněnou kategorii 1. úrovně (viz opět tabulka 4.3.1 na straně 44). Výběr konkrétních titulů novin a časopisů, zastoupení jednotlivých roků vydání i další rozhodnutí tady kromě možností banky závisejí zejména na převládající interpretaci konceptu reprezentativnosti v okamžiku vyvažování korpusu. Důsledkem této neurčitosti jinak velmi podrobné kategorizace je zejména nejednotnost v zastoupení roku vydání publicistických textů v korpusu SYN2000 na jedné straně a SYN2005 spolu se SYN2010 na straně druhé (tento rozdíl je dobře vidět na obr. 4.5.1 na straně 54). V prvním případě převládl přístup založený na recepci, podle níž stav jazyka v roce 2000 přirozeně více ovlivňuje publicistika z roku 1999 než 1990, což se projevuje klesajícím zastoupením textů staršího roku vydání; ve druhém případě naopak vidíme přístup podobný monitorovacím korpusům, v nichž je každý rok vydání zastoupen stejným množstvím textů.

Každý jednotlivý rok vydání je však potřeba také naplnit konkrétními tituly tak, aby i jejich proporce odpovídaly předpokládané čtenosti a zároveň vycházely z možností banky. Tento výběr probíhá prakticky tak, že je pro každý rok vydání, který by měl daný reprezentativní korpus pokrýt, nejdříve sestavena tabulka všech titulů spolu s údajem o množství slov, které jsou pro ně v bance k dispozici. Na základě rozvahy, která bere v úvahu „ideální“ poměry jednotlivých druhů periodik (celostátní/regionální, seriózní/bulvární, novinová/časopisecká) v budoucím korpusu a další podobná kritéria, je pak pro každý titul stanoven koeficient vyjadřující jeho váhu. Ta určuje, jaká část textů daného titulu bude do korpusu skutečně vybrána. Samotný výběr probíhá již automaticky tak, že je vybráno například každé čtvrté číslo (jednotlivá čísla jsou vybírána vždy celá), čímž je zajištěno rovnoměrné zastoupení každého období roku. Výběr jednotlivých titulů je tedy do značné míry subjektivní, také zmíněné „ideální“ poměry jednotlivých druhů periodik jsou při absenci podrobnějších, na recepci založených kritérií stanoveny víceméně ad hoc. Na obr. 4.5.2 na straně 55 je vidět proměnlivost zastoupení jednotlivých publicistických titulů, která je však do značné míry dána také jejich (ne)dostupností v bance.

4.2.7 Lemmatizace a morfologické značkování

Množina textů vybraných z banky v předchozím kroku ještě musí projít řadou procedur, jejichž společným rysem je to, že jsou (nebo byly) průběžně vylepšovány. Jejich výstup je proto považován za přechodný a důsledkem je aplikace těchto procedur až před zveřejněním konkrétního korpusu. Jejich dalším specifikem je to, že jsou prováděny

nástroji, které nebyly vyvinuty v ÚČNK. To platí také o automatické ***tokenizaci*** (rozdělení textu na sled tokenů) a ***segmentaci*** (rozpoznání a označení konců vět), které se spouštějí nejdřív jako nutný předstupeň k vlastní lemmatizaci a morfologickému značkování.

Teprve na ně navazuje lemmatizace spojená s morfologickým značkováním, které již patří k velice sofistikovaným procedurám. Na tomto místě pouze shrnujeme, že jejich cílem je přiřadit každému tokenu v korpusu jeho lemma a morfologickou značku. ***Morfologická značka*** je řetězec, do něhož jsou zakódovány informace o slovním druhu a mnoha dalších morfologických vlastnostech daného tvaru (aktuální a podrobný popis pozičního systému používaného v ČNK lze najít na `http://www.korpus.cz/bonito/znacky.php`). Obojí probíhá současně ve dvou hlavních krocích: v prvním kroku, ***morfologické analýze***, je každému tokenu přiřazena sada všech teoreticky možných morfologických interpretací, tj. párů složených z morfologické značky a jí odpovídajícího lemmatu. Morfologická analýza je založena na rozsáhlém slovníku a probíhá bez ohledu na kontext, takže každým dvěma tokenům téhož typu je přiřazena shodná sada dvojic lemmat a tagů.

Upozorněme zde na velkou homonymii slovních tvarů v češtině, a to jak systémovou – častá shoda tvarů pro nominativ/akuzativ nebo dativ/lokál, pádový synkretismus substantiv vzoru *stavení* atd. –, tak náhodnou – tvary *je* (sloveso vs. zájmeno), *jedli* (sloveso vs. substantivum) a mnoho dalších. Praktickým problémem morfologické analýzy je však také velká homonymie českých slov s cizími, například tvar *an* je v textu málokdy vztažné zájmeno, častěji jde o anglický neurčitý člen, německou předložku, část čínských či jiných exotických jmen, nebo jde prostě o chybu (Křen, 2006a, str. 19). Velkým problémem morfologického slovníku je množství proprií a jejich fluktuace, která je typická zvláště pro publicistiku. Během zpracování každého velkého korpusu je proto slovník morfologické analýzy vždy doplňován o nově nalezená frekventovaná lemmata, jimiž jsou často právě propria.

Případná tvarová homonymie je vyřešena v druhém kroku, ***desambiguaci***, kdy je ke každému tokenu vybrána vždy jedna morfologická interpretace, tedy dvojice lemmatu a morfologické značky z předchozího kroku. Desambiguace probíhá na základě kontextu konkrétního tokenu, v současné době se používá kombinace stochastických a pravidlových metod, které byly podrobně popsány v řadě publikací (zejména Hajič 2004; Petkevič 2006; Spoustová et al. 2007; Jelínek 2008). Celková chybovost desambiguace se u nejlepších algoritmů pohybuje mezi 4–5 %, největším problémem je přitom správné určení jmenného rodu, čísla a pádu v rámci paradigmatu daného lemmatu. Určování slovního druhu – a tedy i s ním úzce související určování lemmatu – má chybovost pouze okolo 0,5 % (Spoustová et al., 2007, str. 73), a je tedy relativně spolehlivé.

Lemmatizace a morfologické značkování jsou poslední kroky, které předcházejí zveřejnění korpusu. Po nich následuje už jen technický převod textů do formátu tzv. vertikálního textu (Rychlý, 2000, str. 29) požadovaného systémem Manatee/Bonito jako vstup pro zaindexování korpusu.

4.3 Reprezentativnost korpusů řady SYN

4.3.1 Kategorizace textů

Reprezentativnost korpusů řady SYN je založena na korpusovém odrazu recepce (nikoli tedy například produkce) psaných textů běžnými uživateli jazyka. Tyto korpusy jsou zároveň vyvážené, tj. představují „ideální“ namíchání jednotlivých variet psaného jazyka tak, aby proporce jejich zastoupení v korpusu závisela v rozhodující míře na jejich čtenosti. Tento koncept předpokládá, že texty s velkým množstvím čtenářů ovlivňují jazyk více než ty, které mají čtenářů méně nebo jsou neveřejné. První takto založené představy o reprezentativnosti byly publikovány již několik let před zveřejněním korpusu SYN2000, prvního korpusu této řady, jako pokus o fundované uchopení nejasné a často podceňované reprezentativnosti (Čermák, 1997; Čermák et al., 1997).

Tabulka 4.3.1 na následujících stranách představuje úplný přehled proporcí zastoupení jednotlivých kategorií v reprezentativních korpusech řady SYN, který – pokud je nám známo – dosud nebyl v takto ucelené podobě publikován. Čísla uvedená pro SYN2005 přitom platí beze změny také pro SYN2010, protože zastoupení jednotlivých kategorií se mezi SYN2005 a SYN2010 nijak nezměnilo. Jde o kombinaci výstupů výzkumů o reprezentativnosti (viz podkapitola 4.3) převedených do dvojic značek *txtype* a *genre*, podle nichž byly korpusy reálně vyvažovány (viz oddíl 4.2.6). Považujeme za potřebné upozornit, že zde udávaná čísla pro korpus SYN2005 se úplně neshodují s čísly, která ve svém článku uvádí Králík (2004, str. 140) a která později doznala změn v beletristické části. V tomtéž článku je také vysvětlena změna postavení administrativních textů, které byly pro SYN2005 „rozpuštěny“ do jednotlivých odborných kategorií, viz také Králík a Šulc (2005, str. 364–365).

Kategorie uvedené v tabulce můžeme rozdělit do tří úrovní uspořádaných v zásadě hierarchicky: na nejvyšší úrovni jde o členění na beletrii, odbornou literaturu a publicistiku, které odpovídá atributu *txtype_group* a které je plně odvoditelné z hodnoty *txtype* (viz oddíl 4.2.4). Kategoriemi druhé úrovně rozumíme všechny beletristické kategorie a souhrnné, tučně vytištěné kategorie v odborné literatuře (např. „vědy o umění“). Všechny ostatní kategorie odborné literatury (např. „hudba“) jsou kategoriemi třetí úrovně. Poznamenejme také, že zatímco pro rozlišení beletristických kategorií jsou často potřeba konkrétní hodnoty *txtype* a *genre*, rozlišení odborných kategorií druhé a třetí úrovně vyplývá pouze z hodnoty *genre* (je-li ovšem odborný také *txtype*, viz ta-

bulka 4.2.1 na straně 37; prázdné políčko *txtype* u nich lze interpretovat jako libovolný *txtype* odpovídající rozlišení pro kategorii vyšší úrovně).

Procenta udávají proporce v množství slov, nikoli pozic (viz oddíl 4.2.1), a jsou vždy vztažena k celému korpusu. To znamená, že například zastoupení kategorie „písně“ se mezi korpusy SYN2000 a SYN2005 vzhledem k beletrii jako celku téměř nezměnilo, protože v obou případech tvoří zhruba její 1 % (neboli 0,16:15 přibližně odpovídá 0,38:40). Procenta tištěná tučně jsou vždy součty proporcí kategorií nižších úrovní uváděné pro větší přehlednost.

txtype	*genre*	kategorie	SYN2000	SYN2005
		BELETRIE	**15,00 %**	**40,00 %**
VER		básně	0,65 %	0,79 %
SON		písně	0,16 %	0,38 %
SCR		dramatické texty, scénáře	0,21 %	0,50 %
NOV		romány	7,65 %	15,25 %
COL		povídky	1,70 %	5,19 %
	CRM	detektivky	1,06 %	4,89 %
	SCF	sci-fi, fantasy	0,42 %	2,37 %
	JUN	literatura pro děti a mládež	0,19 %	2,16 %
FAC	TRV	cestopisy, průvodce	1,04 %	2,64 %
FAC	MEM	biografie, vzpomínky	0,73 %	2,59 %
FAC	CHR	kroniky, deníky	0,46 %	1,53 %
FAC	LET	dopisy	0,37 %	0,72 %
IMA		jiné imaginativní texty	0,36 %	0,99 %
		PUBLICISTIKA	60,00 %	33,00 %
		ODBORNÁ	**25,00 %**	**27,00 %**
ADM		administrativa	0,49 %	
		vědy o umění	**3,48 %**	**2,27 %**
	MUS	hudba	0,72 %	0,33 %
	CIN	film	0,56 %	0,41 %
	TVS	televize	0,25 %	0,25 %
	ARC	architektura	0,34 %	0,31 %

txtype	*genre*	kategorie	SYN2000	SYN2005
	ART	výtvarné umění	0,55 %	0,59 %
	THE	divadlo	0,43 %	0,11 %
	LIT	literární věda	0,63 %	0,25 %
	ARS	jiné		0,02 %
		sociální vědy	**3,67 %**	**4,34 %**
	HIS	historie	0,27 %	1,06 %
	PSY	psychologie	0,07 %	0,92 %
	EDU	pedagogika	0,70 %	0,45 %
	SOC	sociologie	0,16 %	0,66 %
	PHI	filozofie	0,09 %	0,47 %
	INF	knihovnictví	0,22 %	0,13 %
	POL	politologie	1,39 %	0,29 %
	LIN	lingvistika	0,31 %	0,18 %
	ETH	etnografie	0,46 %	0,17 %
	HUM	jiné		0,01 %
		právo a bezpečnost	**0,82 %**	**2,10 %**
	JUR	právo	0,40 %	1,13 %
	MIL	vojenství	0,21 %	0,45 %
	SEC	bezpečnost	0,21 %	0,42 %
	LAW	jiné		0,10 %
		přírodní vědy	**3,37 %**	**4,34 %**
	AGR	zemědělství, lesnictví	1,00 %	0,52 %
	MED	lékařství	1,15 %	0,95 %
	ZOO	zoologie	0,13 %	0,38 %
	BOT	botanika	0,12 %	0,40 %
	BIO	biologie	0,03 %	0,26 %
	ANT	antropologie	0,03 %	0,11 %
	CHE	chemie	0,13 %	0,18 %

txtype	*genre*	kategorie	SYN2000	SYN2005
	MAT	matematika	0,13 %	0,25 %
	LOG	logika	0,03 %	0,09 %
	GGR	geografie	0,05 %	0,19 %
	AST	astronomie	0,03 %	0,15 %
	PHY	fyzika	0,08 %	0,20 %
	MET	meteorologie	0,14 %	0,08 %
	GEO	geologie	0,09 %	0,09 %
	ENV	ekologie	0,23 %	0,47 %
	NAT	jiné		0,02 %
		technika	**4,61 %**	**4,12 %**
	TRA	doprava, komunikace	1,13 %	0,54 %
	ENE	energetika	0,91 %	0,20 %
	IND	průmysl	1,37 %	0,73 %
	COM	počítačová technika, informatika	0,66 %	1,78 %
	BUI	stavebnictví	0,41 %	0,65 %
	STA	standardizace, metrologie	0,13 %	0,13 %
	TEC	jiné		0,09 %
		ekonomie a řízení	**2,27 %**	**2,90 %**
	ECO	ekonomie, obchod	1,16 %	1,55 %
	MAN	management	0,77 %	0,92 %
	MER	zbožíznalství, spotřebitel	0,34 %	0,39 %
	ECN	jiné		0,04 %
		víra	**0,74 %**	**1,18 %**
	REL	náboženství, teologie	0,67 %	0,62 %
	EXC	nadpřirozeno, magie	0,07 %	0,51 %
	BEL	jiné		0,05 %
		životní styl	**5,55 %**	**5,75 %**
	HOU	domácí práce, stravování, byt	0,57 %	1,29 %

txtype	*genre*	kategorie	SYN2000	SYN2005
	SPO	sport	1,18 %	1,21 %
	SCT	společenský život, drby	0,09 %	1,19 %
	AMU	zábava, koníčky, hry	3,39 %	0,94 %
	MIN	minority	0,09 %	0,35 %
	REG	region	0,23 %	0,70 %
	LIF	jiné		0,07 %

Tabulka 4.3.1: Zastoupení typů textu a žánrů v reprezentativních korpusech.

Postup při stanovení konkrétních proporcí jednotlivých kategorií pro korpus SYN2000 vysvětluje Králík (2001). Jde o kombinaci výsledků z mnoha zdrojů různého druhu, od srovnání množství existujících časopisů s určitým odborným zaměřením přes statistická data o výpůjčkách ve veřejných knihovnách až po sociologické průzkumy zjišťující, co a jak často lidé čtou. Jejich výstupem je stanovení ideálního poměru mezi beletrií, odbornou literaturou a publicistikou v korpusu SYN2000, a to včetně dalšího dělení beletrie a odborné literatury na kategorie nižší úrovně a jejich přesného procentuálního zastoupení. Nejsou tu však zmíněny poměry 3. úrovně, tedy v rámci odborné literatury, ačkoli i ony patří mezi výsledky těchto průzkumů. Podrobnou kategorizaci na všech úrovních naopak uvádí Šulc (2001). Je sice bohužel bez konkrétního procentuálního zastoupení, ale zato s uvedením typu textu a žánru, které jsou nezbytné pro mapování těchto kategorií na konkrétní texty v bance.

Jak je patrné z tabulky 4.3.1, došlo ve složení korpusu SYN2005 ve srovnání s korpusem SYN2000 k mnoha dílčím a jedné zásadní změně. Touto zásadní změnou je výrazně vyšší podíl beletrie (nárůst z 15 % na 40 %) a odpovídající pokles podílu publicistiky (ze 60 % na 33 %). Králík (2004) je vysvětluje výsledky dalších dvou průzkumů provedených koncem roku 2001, konkrétně poklesem zájmu o politické dění spojeným s odklonem od čtení novin od roku 1996, kdy se konaly starší průzkumy; ty navíc vyhodnocovaly pouze četnost kontaktu s novinami, na rozdíl od času nad nimi skutečně stráveného v průzkumu novějším. Tento článek věnovaný návrhu aktualizace struktury pro další reprezentativní korpus uvádí opět přesná čísla pouze 1. a 2. úrovně; poměry uvnitř beletrie se ale na základě těchto průzkumů neměnily, čísla byla pouze přepočítána pro její celkově větší zastoupení. Ke změnám naopak došlo uvnitř odborné literatury, ty jsou však vysvětleny pouze povšechným odkazem na jeden z nových průzkumů.

Složení korpusu SYN2010 bylo bez dalších průzkumů převzato z korpusu SYN2005. Jde o řešení pragmatické a do jisté míry pochopitelné, považujeme ho však za nevhodné,

protože vyváženost korpusu SYN2010 je tak z hlediska recepce současného psaného jazyka jen obtížně obhajitelná. Nelze přece předpokládat, že by se recepce psaného jazyka v letech 2005–2009 vůbec nezměnila, zvláště změnila-li se v letech 2000–2004 skutečně radikálně a tato změna byla ve složení korpusu SYN2005 také reflektována. Toto zpochybnění vyváženosti korpusu SYN2010 se ale týká pouze neměnnosti podílů jednotlivých kategorií. Zdůrazněme proto, že shodné složení korpusů SYN2005 a SYN2010 je pro metodu popisovanou v kapitole 5 spíše výhodou a že tato metoda musí vzít v úvahu také rozdílné složení korpusu SYN2000, takže toto zpochybnění neměnnosti podílů jednotlivých kategorií by na výsledky práce nemělo mít prakticky žádný vliv.

Čísla, která jsou výsledky výzkumů, mají z dnešního pohledu dva hlavní nedostatky: v první řadě vůbec nepočítají s novými, ryze internetovými žánry (chaty, blogy, dokonce ani e-maily), což je sice vzhledem k době jejich vzniku pochopitelné, u korpusu SYN2010 však už začíná být problematické. Jako odpověď na tuto námitku můžeme samozřejmě korpusy řady SYN považovat za reprezentaci pouze tištěného jazyka, zužuje se tak ale jejich vypovídací hodnota o psaném jazyce jako celku.

Druhým nedostatkem je zařazení celé kategorie 2. úrovně „životní styl“ do odborné literatury; její velkou část tvoří časopisy typu Instinkt, Story, Playboy, Vlasta, Maminka atd., které v rámci daných kategorií nebylo při anotaci možné zařadit jinam. To je zřejmě důsledek tehdejšího podcenění nárůstu jejich obliby v současné době, kdy se mohlo zdát, že lze vystačit s členěním na časopisy neodborné (a tedy patřící do publicistiky, např. Reflex) a zájmové (a tedy patřící do odborné literatury, např. Podlahy a interiér). Nejistý status všech kategorií patřících pod „životní styl“ je patrný i v tom, že proporce jejich zastoupení doznaly mezi korpusy SYN2000 a SYN2005 z odborných kategorií největších změn.

Zdůrazněme však také dvě výhody této kategorizace, které jsou podle našeho názoru podstatnější než zmíněné dílčí výhrady. Jde jednak o její velkou podrobnost, která napomáhá udržovat pestrost reprezentativních korpusů a zajišťuje celkovou věrnost reprezentace při přechodu na vyšší úrovně. Další velkou výhodou je – zvláště z hlediska srovnávání korpusů – stálost celého systému kategorizace textů, který se od vzniku korpusu SYN2000 téměř nezměnil.

4.3.2 Vymezení synchronie

Další důležitou otázkou, kterou bylo potřeba vyřešit při budování korpusů řady SYN jako reprezentantů synchronní psané složky ČNK, je stanovení hranice mezi synchronií a diachronií. Prakticky vzato jde o určení kritérií pro stanovení „současnosti“ textu, která musejí být splněna, aby text mohl být zařazen do synchronního korpusu. Protože je koncept reprezentativnosti založen na recepci, bere v úvahu především čtenost

jednotlivých textů, tedy míru jejich vlivu na současného čtenáře včetně případných reedic. Je přitom zřejmé, že zatímco starší beletristická díla mohou být stále populární, stejně staré publicistické texty čte jen málokdo; hranici mezi synchronií a diachronní je tedy potřeba vést v závislosti na typu textu. Výsledkem těchto úvah bylo pro korpus SYN2000 stanovení následujících kritérií (Čermák, 2001a, str. 24 a Kučera, 2002, str. 251–252): rok prvního vydání 1990 jako přirozená hranice synchronie pro texty informativní (publicistické a odbornou literaturu); rok 1990 zároveň tvoří i hranici jádra synchronie pro texty imaginativní, ovšem s tím, že toto jádro může být doplněno i texty staršími, pokud se jejich autor narodil po roce 1880 a zároveň byly vydány (ne nutně poprvé) po roce 1945.

Při kontinuálním mapování jazyka představovaném korpusy řady SYN se hranice synchronie musí pochopitelně dříve nebo později začít posouvat. Tyto posuny jsou v publicistice jasně dány tím, že do každého z korpusů byly zařazeny pouze texty z období, které by měl daný korpus pokrývat, tj. 1990–1999 pro SYN2000, 2000–2004 pro SYN2005 a 2005–2009 pro SYN2010. V zastoupení jednotlivých roků vydání v rámci těchto období však došlo k odklonu od pojetí reprezentativnosti založeného na recepci, který je demonstrován změnou složení publicistických textů zmíněnou již v oddílu 4.2.6. Původní pokus o zachycení jazyka publicistiky v jednom konkrétním okamžiku (konkrétně v roce 2000), který starší publicistika ovlivňuje přirozeně méně než novější, byl nahrazen přístupem připomínajícím monitorovací korpusy se stejným zastoupením každého roku vydání; tento posun je vidět na obr. 4.5.1 na straně 54.

Hranice synchronie se v beletrii a odborné literatuře nijak neměnila teoreticky, prakticky však došlo k posunům daným snahou o zařazování co nejnovějších textů do každé z kategorií, jak je zřejmé z tabulek 4.3.2 a 4.3.3 (údaje jsou založeny na korpusu SYN ve verzi z 20. prosince 2010). Tyto posuny jsou vidět mj. na nule u předrevolučních odborných textů v SYN2010, i když by se striktně vzato žádné takové texty neměly vyskytovat ani v ostatních dvou korpusech. Podstatné však je, že se hranice synchronie v odborné literatuře posouvala rychleji než v beletrii, která je tak považována z hlediska udržování si svého vlivu v čase za stálejší. Tato větší stálost beletrie je vidět také na tom, že v jednotlivých korpusech není nejvíce beletristických textů vždy v posledním pětiletém období, ale většinou v období předposledním; v tomto případě však jde spíše o nezamýšlený důsledek stavu banky, což platí zvláště pro korpus SYN2005. Zejména z hlediska diachronní srovnatelnosti je však nepříjemné zahrnování reedic starších textů, tedy nejenom toho, co v daném období skutečně vzniklo. Na tento problém poukazuje také Kučera (2011, str. 71) v kontextu popisu rozdílů mezi synchronní a diachronní složkou ČNK, které brání jejich snadnému propojení. Zahrnování reedic sice přirozeně vychází z recepce, na níž je celý koncept založen, reedice však bohužel nejsou snadno zjistitelné, protože datace vzniku textu ve strukturních atributech chybí. Takto zvýšená setrvačnost v reflektování jazykových změn však na druhé straně

může odpovídat realitě, a proto samo zahrnování reedic do korpusu nelze považovat jednoznačně za nevýhodu.

	do r. 1989	1990–1994	1995–1999	2000–2004	2005–2009
SYN2000	36	186	190	0	0
SYN2005	73	91	394	216	0
SYN2010	14	70	142	304	300

Tabulka 4.3.2: Počty textů v beletrii podle roku vydání.

	do r. 1989	1990–1994	1995–1999	2000–2004	2005–2009
SYN2000	6	52	432	0	0
SYN2005	4	24	253	479	0
SYN2010	0	5	78	154	663

Tabulka 4.3.3: Počty textů v odborné literatuře podle roku vydání.

4.3.3 Shrnutí

Tabulka 4.3.4 shrnuje největší rozdíly ve složení reprezentativních korpusů řady SYN, kterými jsou především výrazně změněné poměry hlavních typů textu (včetně menších korekcí i v jejich rámci), dále přesné chápání toho, co pod který hlavní typ textu zařadit („striktní *txtype*“, viz oddíl 4.2.6) a rozložení roku vydání v publicistice, které přestává odrážet recepci a blíží se spíše pojetí monitorovacích korpusů. Je vidět, že zatímco složení korpusů SYN2005 a SYN2010 je skutečně srovnatelné, korpus SYN2000 se od nich v mnohém odlišuje, zvlášť vezmeme-li v úvahu i to, že na rozdíl od obou novějších pokrývá celých deset let.

	zastoupení beletrie	zastoupení odborné	zastoupení publicistiky	striktní *txtype*	roky vydání v publicistice
SYN2000	15 %	25 %	60 %	NE	1990–1999 graduálně
SYN2005	40 %	27 %	33 %	ANO	2000–2004 rovnoměrně
SYN2010	40 %	27 %	33 %	ANO	2005–2009 rovnoměrně

Tabulka 4.3.4: Hlavní rozdíly mezi reprezentativními korpusy řady SYN.

Rozdíl ve složení SYN2000 na jedné straně a SYN2005 spolu se SYN2010 na straně druhé se tedy projevuje na více rovinách, z čehož je patrné, že nejde o pouhou reflexi rozdílů ve čtenosti jednotlivých hlavních typů textu, ale spíše o koncepční obrat. Jeho dopady na diachronní srovnatelnost těchto korpusů však lze do značné míry eliminovat normalizací a srovnáváním po jednotlivých hlavních typech textu, jak ukážeme dále v kapitole 5.

Závěrem však zdůrazněme, že hlavní výhodou popsaného konceptu reprezentativnosti je především to, že je založen právě na recepci psaného jazyka a podložen mnoha výzkumy. Přes řadu dílčích výhrad, které byly v této podkapitole záměrně zvýrazněny, tak představuje fundované řešení problému vztahu mezi korpusem a jazykovou realitou, který je pro vypovídací hodnotu veškerých na korpusu založených zkoumání zásadní. Jeho spolehlivost dále zvyšuje podrobnost a stálost kategorizace dané tabulkou 4.3.1 na straně 44, jíž se řídil výběr textů do všech korpusů řady SYN, a která se tak může stát relativně spolehlivým vodítkem pro odlišení vlivu složení korpusu od vlivů daných jazykovým vývojem. Při žádné podobné kategorizaci se samozřejmě není možné vyhnout jisté míře subjektivnosti, na nejvyšší úrovni je však celý tento systém spolehlivý, a hlavně dlouhodobě konzistentní.

4.4 Hlavní rysy korpusu SYN

Všechny korpusy řady SYN jsou ***referenční***, tj. zůstávají od svého zveřejnění po celou dobu neměnné, což platí i pro veškerou přidanou informaci a metadata včetně lemmatizace, morfologického značkování a bibliografické anotace. Referenčnost má především tu výhodu, že všechny dotazy, statistiky apod. jsou opakovatelné a dávají stále stejné výsledky. Na druhou stranu má ale i nevýhody, jejichž závažnost se zveřejňováním dalších korpusů řady SYN postupně rostla (Křen, 2009).

Každý korpus řady SYN je totiž zpracován vždy nejnovějšími nástroji, které jsou v době jeho vzniku dispozici, což se týká zejména nástrojů popsaných v oddílu 4.2.7. Ty však prošly od zveřejnění korpusu SYN2000 řadou výrazných vylepšení, které se týkaly především kvality lemmatizace a morfologického značkování: změnil se způsob zpracování některých jazykových jevů, bylo doplněno mnoho nových slovních tvarů do slovníku morfologické analýzy a v neposlední řadě se zvýšila úspěšnost desambiguace včetně oprav mnoha hrubých chyb.[1] Vylepšení ve zpracování jednotlivých korpusů se však projevila i na úrovni slovních tvarů: v korpusu SYN2000 tak například najdeme složeniny typu *červeno-černý* rozdělené na tři tokeny (*červeno - černý*), zatímco ve

[1] V korpusu SYN2000 desambiguace často vybírala velice nepravděpodobné interpretace některých slovních tvarů, například slovní tvar *mračen* je v něm v 75 % případů označkován jako pasivní participium lemmatu *mračit*, při zpracování korpusu SYN2005 byl zase nevhodně použit tzv. guesser odhadující slovní druh a další morfologické kategorie neznámých slov přímo ze slovních tvarů atd.

všech novějších korpusech zůstaly zachovány jako token jediný (*červeno-černý*), což je pouze důsledkem jiného způsobu tokenizace.

Nevyhnutelným důsledkem referenčnosti korpusů řady SYN je však nemožnost všechny tyto úpravy a vylepšení do již zveřejněných korpusů promítnout. Zejména zpracování korpusu SYN2000 je tak již dlouho nevyhovující, protože odráží stav nástrojů na zpracování jazyka v době před více než deseti lety. To kromě postupného zastarávání způsobu zpracování referenčních korpusů vede také k tomu, že každý z nich je zpracován odlišným způsobem. Tento fakt má – zvláště spolu s odlišným složením reprezentativních korpusů řady SYN popsaným v podkapitole 4.3 – vážné negativní dopady na srovnatelnost a znesnadňuje tak jejich použití pro sledování jazykového vývoje: i v případě nalezení podstatných rozdílů ve frekvencích slovních tvarů nebo lemmat je pro běžné uživatele obtížné zjistit příčinu, a oddělit tak rozdíly způsobené pouhým zpracováním od rozdílů daných složením jednotlivých korpusů (které by v případě reprezentativních korpusů měly odpovídat změnám v jazyce). Na stránce `http://www.korpus.cz/srovnani.php` proto byly jako další zdroj dat zveřejněny Srovnávací frekvenční seznamy z korpusů SYN2000 a SYN2005, které uživatelům umožňují srovnávání lexikálních frekvencí na základě jednotně zpracovaných korpusů; koncem roku 2010 se na `http://www.korpus.cz/srovnani10.php` objevila jejich aktualizovaná verze zahrnující i korpus SYN2010.

Koncepčním řešením celé situace se však stal až korpus SYN – spojení všech korpusů řady SYN jednotně zpracovaných nejnovějšími verzemi dostupných nástrojů (Křen, 2009). Jeho velikost je díky disjunktnosti všech korpusů této řady daná součtem velikostí jednotlivých korpusů (až na malé rozdíly způsobené tokenizací), což v současné době znamená 1,3 miliardy slov. Korpus SYN samozřejmě není reprezentativní a je záměrně nereferenční, protože se v budoucnu počítá s jeho dalšími aktualizacemi. Kromě výše zmíněných vylepšení ve zpracování textů přitom korpus SYN reflektuje i doplňky, opravy a sjednocení bibliografické informace dodávané k textům při anotaci (viz oddíl 4.2.4). Dále je z hodnot strukturního atributu *syn* možné poznat, ze kterého referenčního korpusu řady SYN daný text pochází, a tvořit podle něj subkorpusy odpovídající svým složením původním korpusům.

Korpus SYN je tedy možné chápat jako v mnoha ohledech průběžně aktualizovaný „obal“ jednotlivých korpusů řady SYN umožňující práci s nejnovějším způsobem zpracovanými verzemi všech textů zařazených do původních referenčních korpusů. Jeho jednotné zpracování navíc umožňuje snadnou srovnatelnost veškerých dat včetně lemmatizace a morfologického značkování. Kromě zpětné kompatibility tak vlastně není důvod používat původní referenční korpusy, a proto jsou také v této práci všechny texty převzaty právě z korpusu SYN.

4.5 Subkorpusy a způsob jejich výběru

Tato podkapitola podrobně popisuje několik skupin subkorpusů korpusu SYN, na nichž je založena celá další práce. Jejich volba vyplývá z diskuse použité metody popsané v kapitole 5. Tyto subkorpusy jsou neveřejné a vznikly díky možnosti přímého přístupu k datům. Byly vybrány perlovským skriptem přímo z tzv. vertikálního textu korpusu SYN ve verzi zveřejněné 20. prosince 2010 a následně zaindexovány (viz oddíl 4.2.7). Vznikly tak dvě podoby každého subkorpusu, každá pro jiné účely: vertikální text jako vstup pro statistické procedury popsané v kapitole 5 a zaindexovaná podoba pro přístup pomocí modulu, který pro dotazování používá dotazovací jazyk CQL korpusového manažeru Manatee. To umožňuje využít možností celého dotazovacího aparátu, pomocí tohoto modulu byly vygenerovány průběhové grafy používané v kapitole 6.

Celkem bylo vytvořeno 48 překrývajících se subkorpusů rozdělených do 4 skupin:

- repre_KKKK, kde $KKKK \in \{2000, 2005, 2010\}$, odpovídající celým reprezentativním korpusům SYN2000, SYN2005 a SYN2010 (3 subkorpusy); např. repre_2005 je subkorpus vybraný z korpusu SYN tak, aby jeho složení odpovídalo původnímu reprezentativnímu korpusu SYN2005;

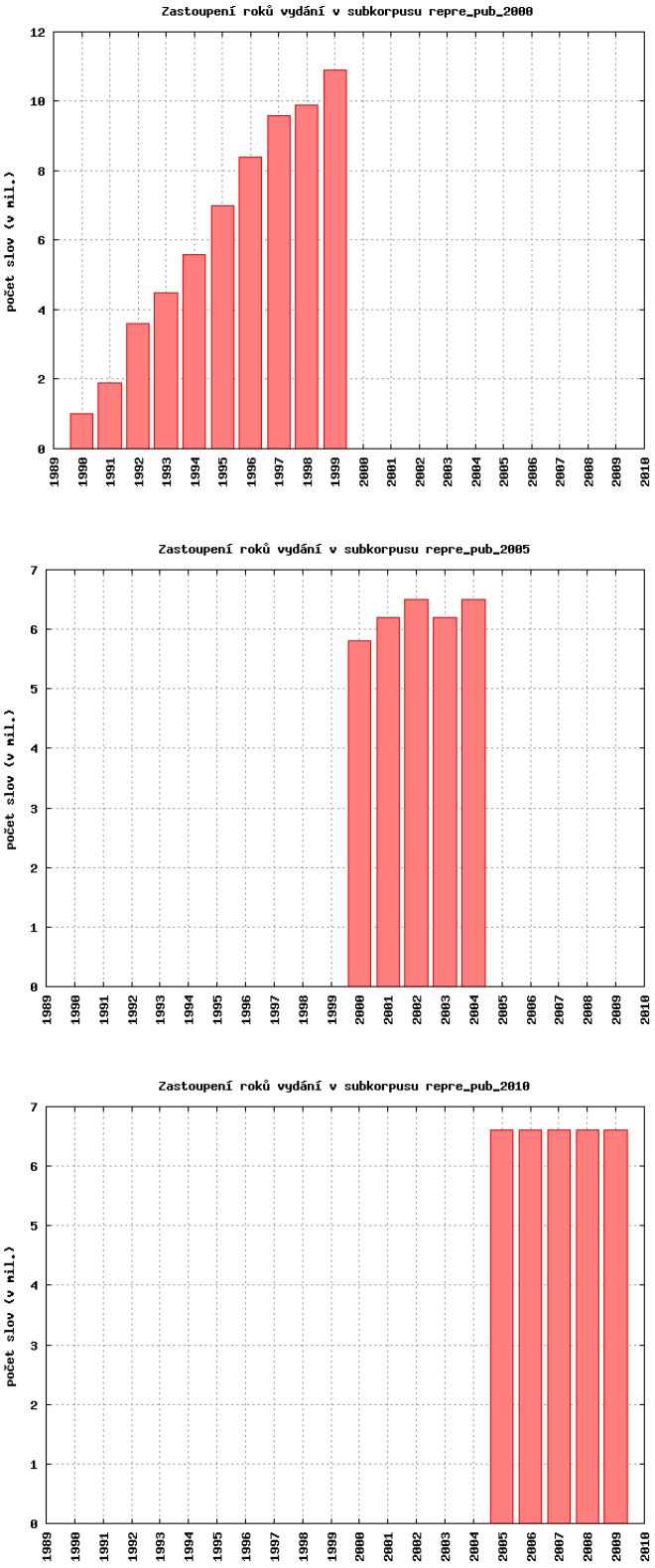
- repre_TTT_KKKK, kde $TTT \in \{bel, odb, pub\}$ a $KKKK \in \{2000, 2005, 2010\}$, odpovídající částem subkorpusů repre_KKKK dále členěným podle hodnoty atributu *txtype_group* na beletrii, odbornou literaturu a publicistiku (9 subkorpusů); např. repre_bel_2005 je subkorpus vybraný z korpusu SYN tak, aby jeho složení odpovídalo beletristické části původního reprezentativního korpusu SYN2005;

- pub_RRRR, kde $RRRR \in \{1992, 1993, ..., 2009\}$, odpovídající veškeré publicistice daného roku vydání v korpusu SYN rozlišené podle hodnoty atributu *txtype_group* (18 subkorpusů); např. pub_2006 je subkorpus obsahující veškerou publicistiku z korpusu SYN vydanou v roce 2006;

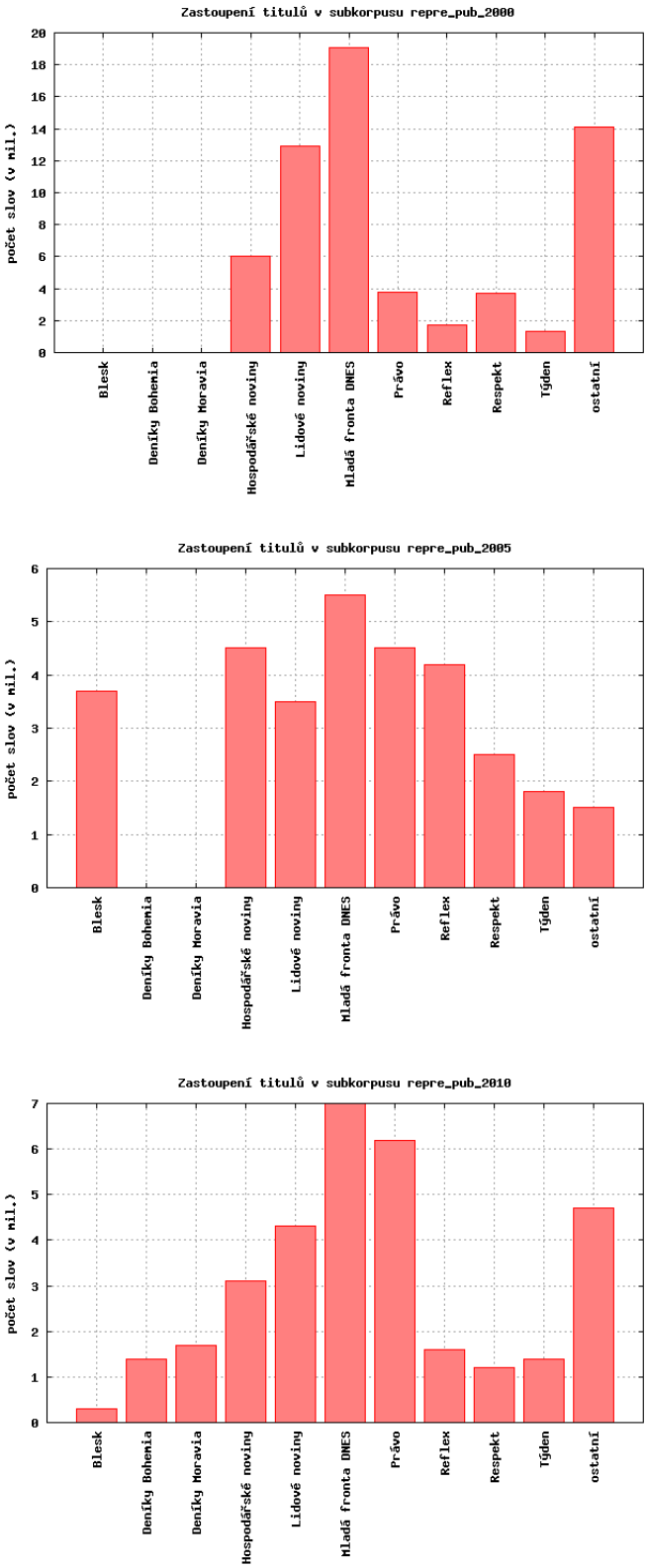
- mf_RRRR, kde $RRRR \in \{1992, 1993, ..., 2009\}$, odpovídající veškeré Mladé frontě DNES (MFD) daného roku vydání v korpusu SYN rozlišené podle identifikátoru opusu (hodnota atributu *id* splňující regulární výraz `mf[0-9]{6}`; 18 subkorpusů); protože každé vydání MFD má *txtype=PUB*, je každý subkorpus mf_RRRR podmnožinou odpovídajícího pub_RRRR, např. mf_2001 je subkorpus obsahující veškerou MFD z korpusu SYN vydanou v roce 2001.[2]

Na následujících stranách uvádíme řadu grafů znázorňujících nejdůležitější parametry složení těchto subkorpusů. Na obr. 4.5.1 je vidět již zmíněný rozdíl v zastoupení jednotlivých roků vydání publicistických textů v reprezentativních korpusech řady

[2]Toto tvrzení neplatí pro mf_2005 až mf_2009, jak vysvětlíme dále.



Obrázek 4.5.1: Zastoupení jednotlivých roků vydání v subkorpusech repre_pub_KKKK.



Obrázek 4.5.2: Zastoupení jednotlivých titulů v subkorpusech repre_pub_KKKK.

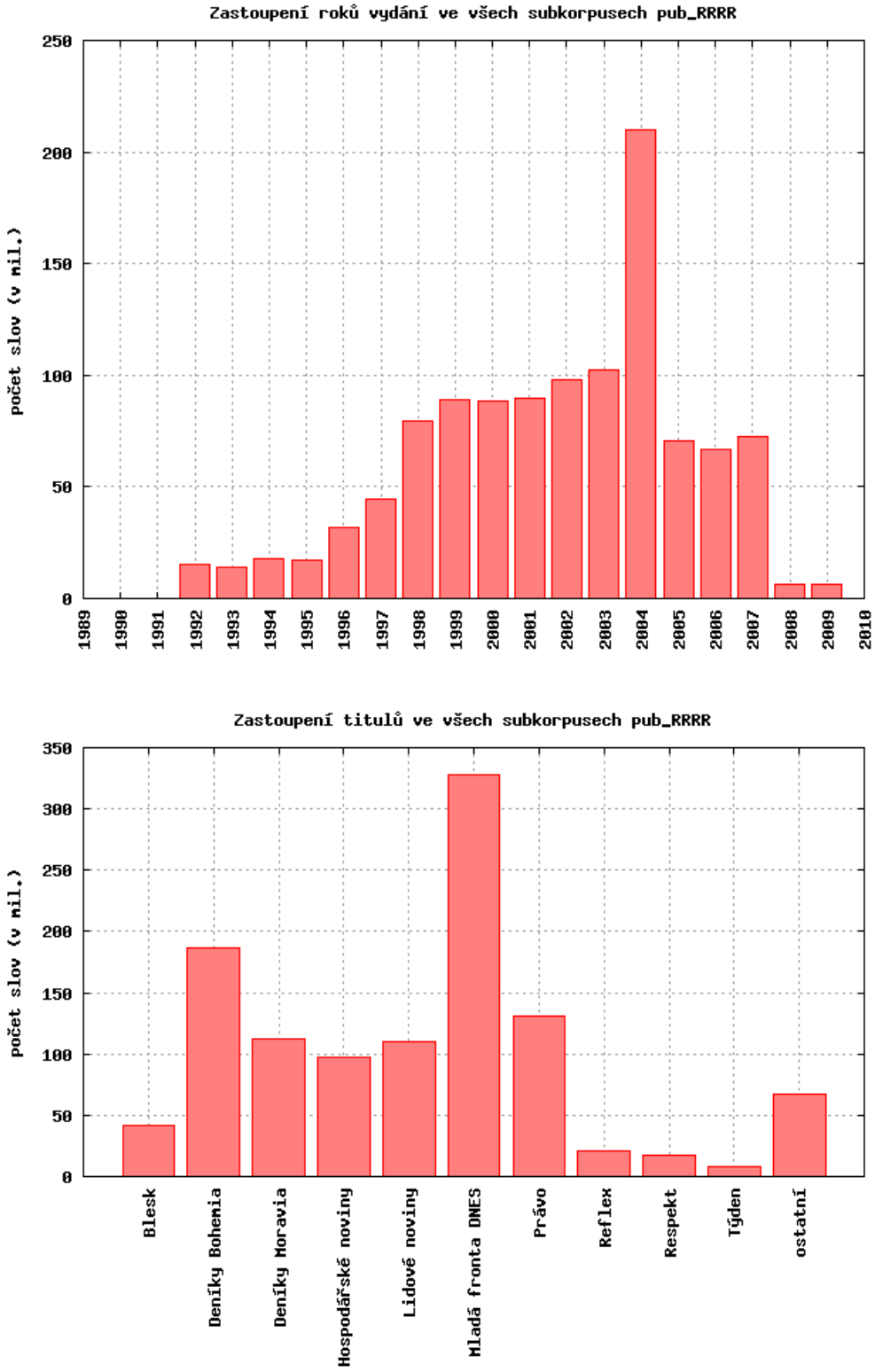
SYN, který se samozřejmě odrazil také ve složení subkorpusů repre_pub_KKKK. Dodejme, že kolísání celkového počtu slov v subkorpusu repre_pub_2005 je dáno jinou verzí tokenizace a sjednocováním kategorizace některých odborných periodik (a s ním spojeným „přesunem“ těchto textů do odborné literatury), k nimž došlo po zveřejnění korpusu SYN2005, a které se tak projevilo až v korpusu SYN. V původní, referenční verzi korpusu SYN2005 je každý rok vydání zastoupen (přibližně) stejným počtem slov, podobně jako je tomu u repre_pub_2010.

Obr. 4.5.2 ukazuje proměny složení reprezentativních korpusů v zastoupení jednotlivých publicistických titulů. Jak již bylo uvedeno v oddílu 4.2.6, je jejich složení dáno především dostupností potřebného množství jednotlivých titulů v bance, a vysvětluje tak například absenci Blesku v korpusu SYN2000. Složení publicistiky v těchto korpusech je tedy velmi nepravidelné a není přitom podepřeno žádnými výzkumy ani jinými čísly, která by odrážela čtenost. Do jisté míry však reflektuje nárůst vlivu produkce vydavatelství Vltava-Labe-Press (VLP), jehož regionální tituly jsou tady soustředěny do dvou skupin, české a moravské (Deníky Bohemia a Deníky Moravia). Grafy pro jednotlivé subkorpusy ukazují počty slov pro vždy týchž deset titulů s největším zastoupením v publicistice korpusu SYN jako celku.

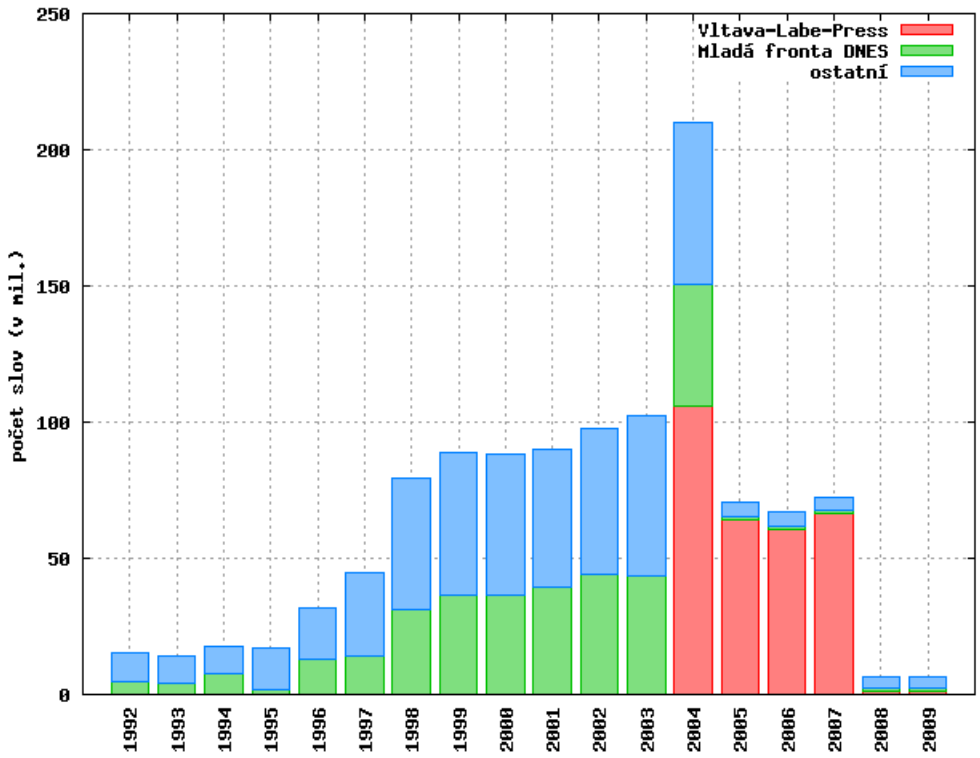
Obr. 4.5.3 ukazuje počty slov pro jednotlivé roky vydání a tituly ve sjednocení všech subkorpusů pub_RRRR, tedy ve veškeré publicistice korpusu SYN kromě roků vydání 1989, 1990 a 1991. Publicistika z těchto let je totiž v datech korpusu SYN zastoupená velice sporadicky (jedinými tituly jsou Informační servis, Respekt a část ročníku 1991 Lidových novin) a žádná další není k dispozici ani v bance; proto byla spodní hranice pro řadu pub_RRRR stanovena právě na rok 1992.

Vzhledem k tomu, že velká většina publicistických textů v korpusu SYN pochází z korpusů SYN2006PUB a SYN2009PUB, má jejich složení rozhodující vliv i na složení subkorpusů pub_RRRR. Tento vliv se projevuje jednak v malém (byť reprezentativním) zastoupení textů z let 2008 a 2009, jejichž zdrojem pro korpus SYN je pouze korpus SYN2010, a také v na první pohled nápadné špičce v roce 2004. Ta se vyskytuje už v korpusu SYN2009PUB (viz `http://www.korpus.cz/syn2009pub.php`) a je nezamýšleným důsledkem toho, že velice rozsáhlá produkce VLP je v bance k dispozici až od začátku roku 2004 a že právě tento ročník byl do korpusu SYN2009PUB zařazen celý. Také všechny ostatní tituly do roku vydání 2004 včetně byly do korpusu SYN2009PUB zařazeny celé,[3] takže subkorpusy pub_RRRR věrně odrážejí stav banky do roku vydání 2004 včetně. Od roku 2005 dále však v korpusu SYN2009PUB najdeme pouze VLP, a i když nejde o kompletní ročníky, nemůže toto množství textů publicistika z korpusu SYN2010 dostatečně kompenzovat. Důsledkem je výrazná převaha VLP v subkorpusech pub_2005, pub_2006 a pub_2007 (viz obr. 4.5.4), která je nejproblematičtějším

[3]Přesněji řečeno do něj byla zařazena všechna jejich čísla vydaná do roku 2004, která byla k dispozici v bance a nebyla přitom zařazena do některého z předchozích korpusů řady SYN.



Obrázek 4.5.3: Celkové složení subkorpusů pub_RRRR.



Obrázek 4.5.4: Zastoupení MFD a VLP ve všech subkorpusech pub_RRRR.

rysem řady pub_RRRR. Velká část produkce VLP je tak koncentrována do těchto tří subkorpusů bez protiváhy v jiných titulech, jejichž zdrojem pro tyto subkorpusy je pouze korpus SYN2010.

Protože subkorpusy pub_RRRR nelze považovat za reprezentativní zástupce publicistiky, byl z korpusu SYN vybrán jeden konkrétní titul, na němž lze vývojové tendence v jazyce zkoumat sice v omezené míře, zato však na základě jasně daných a relativně homogenních dat. Tímto titulem se stala Mladá fronta DNES, a to ze dvou hlavních důvodů: jednak je ze všech celostátních deníků v korpusu SYN nejrozsáhlejší, jednak je její zastoupení dostatečně rovnoměrné. Počty čísel jednotlivých ročníků MFD ve výsledných subkorpusech mf_RRRR jsou vidět na obr. 4.5.5, nepravidelný rozsah z let 1992–1994 následuje propad v roce 1995 (pouze lednová a únorová čísla) a po něm kompletní ročníky 1996–2009 (s výjimkou chybějících prosincových čísel v roce 1997). Tyto výpadky jsou dány chybějícími čísly MFD v bance, kompletní ročník deníku by měl mít okolo 300 čísel. Celkově velký rozsah MFD je dán mj. také zastoupením všech jejích regionálních mutací, zdůrazněme však, že veškeré duplicity mezi nimi by měly být díky použité detekci a odstraňování duplikátů odstraněny (viz oddíl 4.2.5). Ze srovnání obou grafů je také vidět výrazné zvětšení rozsahu průměrného čísla MFD z původních 20 000 – 25 000 slov v letech 1992–1994 na současných zhruba 130 000 slov, jeho skokové zvětšení po roce 1997 odpovídá právě rostoucímu podílu regionálních mutací.

Obrázek 4.5.5: Složení jednotlivých subkorpusů mf_RRRR.

Jednotlivá čísla MFD by neměla obsahovat magazíny; obecně totiž platí, že magazíny vycházející v rámci jednotlivých periodik jsou – pokud je to možné z povahy dat na vstupu konverzí (viz oddíl 4.2.3) – odděleny a kategorizovány zvlášť. Magazín Ona DNES tedy například v korpusu SYN tvoří vždy samostatný opus (s prefixem *mfon*), a to už od začátku jejího vydávání v roce 2006. Jednou z výjimek je Magazín DNES (opusy s prefixem *mfmg*), pro který toto pravidlo platí pouze od roku 2005 včetně. Starší čísla MFD totiž tento magazín mohou, ale nemusejí obsahovat; spolehlivé informace o tom bohužel nejsou k dispozici, pozdější oddělení Magazínu DNES od MFD bylo umožněno změnou způsobu dodávání dat do ÚČNK. Znamená to tedy, že o subkorpusech mf_1993 (kdy se začal Magazín DNES vydávat) až mf_2004 nelze spolehlivě říci, zda Magazín DNES obsahují, či nikoli.

Důležitá poznámka na závěr se týká doplnění novějších ročníků MFD do subkorpusů řady mf_RRRR. Zatímco v korpusu SYN jsou všechny ročníky MFD do roku 2004 včetně pokud možno co nejúplnější (jsou v nich tedy zahrnuta všechna čísla, která jsou k dispozici v bance), z každého z ročníků 2005 až 2009 v něm najdeme jen asi 10 čísel. To je dáno tím, že MFD z těchto let nebyla zařazena do žádného publicistického korpusu (SYN2006PUB ani SYN2009PUB), takže jejím zdrojem v korpusu SYN je pouze SYN2010, v němž je její rozsah pochopitelně omezený, ačkoli v bance je každý z těchto ročníků kompletní. Subkorpusy mf_2005 až mf_2009 proto byly nakonec doplněny o všechna zbývající čísla tak, aby obsahovaly vždy celý ročník MFD.[4] Subkorpusy mf_2005 až mf_2009 tedy jako jediné z používaných subkorpusů netvoří podmnožinu korpusu SYN (ani podmnožinu příslušných subkorpusů pub_RRRR, které takto doplněny nebyly), shodují se s ním však ve způsobu zpracování, lemmatizaci, morfologickém značkování atd., a jsou tedy z tohoto hlediska plně srovnatelné jak s ostatními subkorpusy řady mf_RRRR, tak s celým korpusem SYN.

[4] Za dodatečnou lemmatizaci a morfologické označkování dat MFD z let 2005–2009 děkuji Tomáši Jelínkovi.

5 Popis použité metody

5.1 Úvod

Jak již bylo uvedeno v úvodu práce, je jejím hlavním cílem popsat možnosti a meze detekce tendencí jazykového vývoje metodou založenou na diachronním srovnání synchronních korpusů zachycujících jazyk velice blízkých časových období. Jedním z vedlejších výstupů by pak mělo být také poskytnutí zpětné vazby pro další tvorbu synchronních psaných korpusů řady SYN spojené s případnou korekcí dosavadního pohledu na reprezentativnost synchronních psaných korpusů.

Metoda popsaná v této kapitole a následný rozbor jejích výsledků tvoří hlavní část práce. Metodologicky jde o corpus-driven přístup pracující na dvou úrovních, zabývá se totiž nejenom srovnáváním lexikálních frekvencí, ale i srovnáváním frekvencí lexikálních kombinací. Protože frekvenční rozdíly v typických kombinacích mohou ukazovat na posuny významu, významy nové nebo naopak zastarávající, dotýká se druhá úroveň kromě syntagmatiky také sémantiky. Volba těchto dvou úrovní je dána především pragmaticky, a to dostupností a spolehlivostí lemmatizace a morfologického značkování. Podobný corpus-driven přístup proto v současné době nelze použít na syntax nebo stylistiku přímo, tyto prvky se však mohou objevit při podrobnějším studiu jevů na výše uvedených úrovních.

Může se zdát paradoxní, že volíme corpus-driven přístup v situaci, kdy očekáváme zvýšený vliv složení korpusu na výstupy popisované metody, a kdy se tedy budeme dotýkat vztahu mezi korpusem a jazykovou realitou, který nelze nijak jednoznačně ani objektivně kvantifikovat. Corpus-driven přístup však považujeme za vhodný právě proto, že by měl vynést do popředí všechny jevy, které jsou (na dané jazykové úrovni) významně rozdílné, bez ohledu na naše apriorní očekávání. Objektivnější přístup typu „bottom-up“ proto obhajují i Hilpert a Gries (2009, str. 397): „Frequency developments in temporally ordered corpora often present the analyst with ambiguous, unclear, or otherwise messy data. Interpreting such data on subjective grounds alone can be very problematic and may lead to incongruous conclusions.“ Je ovšem zřejmé, že výsledky corpus-driven metod je potřeba důkladně analyzovat, kapitola 6 je proto věnována rozboru zjištěných rozdílů, jejich typologii a interpretaci příčin.

5.2 Předchozí práce

5.2.1 Nenáhodná povaha jazyka

Metodám a výsledkům popisovaným v této kapitole předcházelo několik studií inspirovaných původně Srovnávacími frekvenčními seznamy (SFS). První verze SFS byla zveřejněna na webových stránkách ÚČNK už v roce 2006 s cílem umožnit veřejnosti fundované srovnávání lexikálních frekvencí mezi korpusy SYN2000 a SYN2005. Problémy se srovnatelností reprezentativních korpusů řady SYN již byly popisovány v podkapitolách 4.3 a 4.4, na tomto místě pouze shrňme, že řada výrazných rozdílů ve složení obou korpusů a v nástrojích použitých na jejich zpracování výrazně znesnadňovala běžným uživatelům interpretaci zjištěných frekvenčních rozdílů. Ukázalo se totiž, že prosté frekvence jednotlivých slov (poněkud vágní, avšak přirozený pojem ***slovo*** budeme v dalším textu používat jako souhrnné označení slovních tvarů a lemmat) v celých reprezentativních korpusech řady SYN se mohou výrazně lišit, zvláště jsou-li typické pro jeden hlavní typ textu, a že tyto rozdíly jsou v rozhodující míře způsobeny právě použitými korpusy. Totéž přitom platí i o celých gramatických kategoriích, jakými jsou například osobní zájmena, jejichž používání je typické pro beletrii. Všechna osobní zájmena můžeme najít dotazem `[tag="P[HP5].*"]`, který v SYN2000 dává 1 362 639 výskytů, zatímco v SYN2005 je jich 2 062 409. Tento více než 50% nárůst je však především důsledkem odlišného složení obou korpusů (15% podíl beletrie v SYN2000 oproti jejímu 40% podílu v SYN2005), a sám o sobě tedy o případných vývojových tendencích v jazyce nevypovídá vůbec nic.

Nutným předpokladem pro sestavení SFS bylo opětovné zpracování obou korpusů nejnovějšími verzemi všech nástrojů, čímž byla zajištěna kompatibilita zpracování. Dále bylo potřeba zajistit také srovnatelnost udávaných frekvencí, pro které se (mezi obecně různě velkými korpusy) běžně používá jejich relativizace nebo také normalizace. ***Normalizovaná frekvence*** udává počet výskytů daného slova v korpusu relativně vzhledem k jeho velikosti, typicky ve výskytech na 1 milion slov (i.p.m., instances per million tokens), a je dána vztahem

$$f_{norm}{=}f \cdot \frac{1000000}{N}$$

kde f je frekvence daného slova v korpusu velikosti N. V SFS však byla namísto toho použita přepočítaná frekvence, a to celková (pro celý korpus) a parciální (pro každý hlavní typ textu zvlášť). ***Celková přepočítaná frekvence*** je dána součtem tří parciálních. Parciální přepočítaná frekvence je definována jako

$$f_{parc} = f \cdot \frac{100000000}{N} \cdot \frac{1}{3}$$

kde f je frekvence daného slova v daném hlavním typu textu velikosti N. Parciální přepočítaná frekvence tedy je frekvence normalizovaná na $\frac{1}{3} \cdot 100000000$ slov. Součet parciálních přepočítaných frekvencí, tj. celková přepočítaná frekvence, však již neodpovídá normalizované frekvenci v žádném reálném korpusu, ale v hypotetickém srovnávacím korpusu, v němž mají všechny hlavní typy textu stejné (tj. třetinové) zastoupení. Celkové přepočítané frekvence jednotlivých slov jsou mezi korpusy srovnatelné stejně jako frekvence normalizované, snaží se však navíc vyjádřit základním způsobem běžnost slova v psaném jazyce tím, že berou v úvahu jeho normalizované frekvence ve všech hlavních typech textu, přičemž každému z nich přikládají stejnou váhu. To samozřejmě neodpovídá výsledkům výzkumů recepce psaného jazyka, na druhou stranu však bylo nutné se pro nějaký poměr rozhodnout. Přesný popis výpočtu přepočítaných frekvencí včetně příkladů použití a interpretace SFS je v jejich nejnovější verzi možné najít na stránce `http://www.korpus.cz/srovnani10.php` a popisuje ho také Křen (2006c).

Na tento článek navazuje další (Křen, 2007), který je pokusem o corpus-driven přístup k vyhodnocování SFS. Pro každé slovo v něm byla významnost rozdílů mezi celkovými přepočítanými frekvencemi v korpusech SYN2000 a SYN2005 vyhodnocena několika různými statistickými mírami ve snaze automaticky najít rozdíly, které jsou statisticky nejvýznamnější. Konkrétně byly použity obecně známé míry log-likelihood (LL), χ^2 a dále nově definovaná CBF daná vztahem

$$CBF = \frac{\chi^2}{\sqrt{f_1 + f_2}}$$

kde f_1 je (celková přepočítaná) frekvence daného slova v korpusu SYN2000 a f_2 je (celková přepočítaná) frekvence téhož slova v korpusu SYN2005. CBF neukazuje žádnou statistickou významnost, jmenovatel ve vzorci je pouhou empiricky stanovenou korekcí χ^2, který bez jejího použití příliš favorizuje frekventovaná slova (viz dále). Všechna slova tedy byla setříděna vždy podle hodnoty dané míry a výsledné seznamy byly nakonec vyhodnoceny. V diskusi se objevilo mnoho otázek a také se ukázalo několik sporných bodů, které nyní ve stručnosti shrneme.

V první řadě vzniká při každém podobném srovnání teoretický problém, které frekvenční rozdíly bychom měli považovat za významnější než jiné, a tedy jak by měly být zjištěné frekvenční rozdíly použitými statistickými mírami v ideálním případě uspořádány. Křen (2007, str. 113) uvádí jako příklad tabulku 5.2.1 s celkovými přepočítanými frekvencemi vybraných lemmat v obou srovnávaných korpusech.

Není totiž zřejmé, zda bychom vzhledem k poměru 0:217 měli za významnější rozdíl považovat nárůst frekvence lemmatu *esemeska*, nebo spíše lemmatu *euro* (nárůst „pouze“ osmiapůlnásobný, ovšem v mnohem vyšší frekvenční hladině), nebo dokonce lemmatu *kraj* (nárůst ani ne trojnásobný, ale s ještě vyšší frekvencí zaručující nenáhodnost dosaženého výsledku). Protože se lze jen těžko shodnout na jediném, „ideálním“

	SYN2000	SYN2005
esemeska	0	217
euro	1 128	9 530
kraj	8 920	24 434

Tabulka 5.2.1: Celkové přepočítané frekvence vybraných lemmat.

způsobu, jak uspořádat desítky až stovky tisíc zjištěných frekvenčních rozdílů podle významnosti, nelze úlohu nalézt vhodnou statistickou míru formulovat jako nalezení takové míry, která by se tomuto ideálnímu uspořádání přiblížila co nejvíce. Přestože tedy výsledky ukazují na někdy zásadní rozdílnost uspořádání frekvenčních rozdílů jednotlivými mírami, kdy může každá z nich označovat za významné rozdíly jiného druhu, je tato pluralita otázkou vhodnosti volby konkrétní míry pro daný účel.

Dalším důležitým bodem je nenáhodná povaha jazyka. Je zřejmé, že při srovnávání lexikálních frekvencí mezi dvěma texty dojdeme k řadě rozdílů daných autorem, stylem, žánrem atd. Podstatné však je, že tyto rozdíly nemají tendenci se vyrovnávat s růstem rozsahu těchto textů (korpusů), a to ani v případě, že jde o jazyk stejného typu: „While it might seem plausible that oddities would in some way balance out to give a population that was indistinguishable from one where the individual words (as opposed to the texts) had been randomly selected, this turns out not to be the case.“ (Kilgarriff, 2001, str. 236). Při velkých rozsazích současných korpusů, a tedy vysokých frekvencích slov v nich obsažených, tak vždy zůstává alespoň malý frekvenční rozdíl, který je považován za statisticky významný.

Příkladem statisticky dobře podložené míry, která z tohoto pohledu příliš favorizuje frekventované jevy, je χ^2 – standardní statistický test, který se pro ověřování nezávislosti jevů používá i v matematické lingvistice. Jeho hodnoty jsou tabelovány, pro každou hodnotu χ^2 a daný stupeň volnosti lze určit hladinu významnosti, tedy pravděpodobnost, s níž je možné pozorovat daný (nebo větší) rozdíl za předpokladu, že je pravdivá nulová hypotéza. V případě prováděného srovnávání lexikálních frekvencí ve dvou korpusech to znamená pravděpodobnost, že pozorovaný frekvenční rozdíl je pouze výsledkem náhodného výběru vzorků z téže populace. Je-li nulová hypotéza zamítnuta, jsou obě populace (velice pravděpodobně) různé, a tedy by (teoreticky) mělo jít o rozdíl způsobený jazykovým vývojem. Nulová hypotéza bývá obvykle zamítnuta na hladině významnosti 0,05, 0,01 nebo dokonce 0,001. Zamítnutí nulové hypotézy na určité hladině významnosti přitom neznamená, že by nulová hypotéza neplatila; znamená to pouze, že je její platnost velice nepravděpodobná. Pokud nulová hypotéza na určité hladině významnosti naopak zamítnuta není, neznamená to, že platí; znamená to pouze, že pro toto zamítnutí není dostatečný empirický základ. V tomto kontextu tedy vysoká hodnota χ^2 nebo podobné míry sice potvrzují, že dva vzorky s největší pravděpodobností

nepocházejí ze stejné populace, nemůže ale odpovědět na otázku, proč tomu tak je. Příčin může být více, při diachronním srovnání je rozdíl v ideálním případě způsoben pouze tím, že mezi jednotlivými obdobími nastala nějaká změna v jazyce, stejně tak je ale možné, že se nám nepodařilo vybrat ve všech ohledech odpovídající vzorky.

Hofland a Johansson (1982) použili χ^2 k nalezení významných rozdílů mezi britskou a americkou angličtinou při srovnávání seznamů slov vygenerovaných ze dvou srovnatelných korpusů, LOB a Brown (viz oddíl 3.3.4). Jako slova s významnými frekvenčními rozdíly přitom byla označena většina frekventovaných funkčních slov. Kilgarriff (2001) ukazuje, že podobných výsledků lze dosáhnout také při srovnávání korpusů stejné jazykové variety, rozdíly zjištěné srovnáváním korpusů LOB a Brown tedy nelze interpretovat jako rozdíly mezi britskou a americkou angličtinou ani nelze na jejich základě zpochybnit srovnatelnost těchto korpusů.

Oakes (1998, str. 28–29) vysvětluje, že hodnota χ^2 roste pro všechny nenáhodně vybrané vzorky spolu s frekvencí, a že tedy i malé frekvenční rozdíly jsou u frekventovaných slov považovány za významné. χ^2 se však ve výše zmíněných (i mnoha jiných) případech používá k testování nulové hypotézy, zda byly srovnávané korpusy vybrány náhodně z větší populace (jazyka). Problém je právě v implicitním předpokladu, že volba slov v textech je náhodným výběrem slov z jazyka. Protože však víme, že jazyk náhodný není, tak vlastně také víme, že takto formulovaná nulová hypotéza neplatí, a χ^2 nám tedy jenom říká, zda už máme dostatek dat k tomu, abychom ji mohli vyvrátit na dané hladině významnosti: „Since words in a text are not random, we know that our corpora are not randomly generated. The only question, then, is whether there is enough evidence to say that they are not, with confidence. In general, where a word is more common, there is more evidence. This is why a higher proportion of common words than of rare ones defeat the null hypothesis.“ (Kilgarriff, 2001, str. 102)

Přestože však nemůžeme říci, zda jsme prokázali (ne)náhodnost daného jevu na určité rovině pravděpodobnosti, zůstává tento problém spíše v teoretické rovině. Výsledky statistických měr lze totiž dále korigovat či jinak kombinovat, prakticky navíc záleží především na tom, jaké druhy výsledků (ve smyslu tabulky 5.2.1) chceme dostat a zda je skutečně dostáváme. Jinými slovy nejde primárně o určení statistické významnosti zjištěných frekvenčních rozdílů, ale především o nalezení vhodného způsobu (či spíše způsobů) jejich uspořádání (ranking). Statistická významnost měřená na korpusech sama o sobě nemusí indikovat rozdíl ve skutečném úzu už vzhledem k problematickému vztahu jazyka a korpusu jako jeho vzorku, takže fakt, že některé míry (např. CBF) nejsou přímo statisticky interpretovatelné, bychom neměli považovat a priori za jejich nevýhodu. Toto tvrzení budeme v dalším textu používat jako zdůvodnění některých empirických úprav standardních matematických postupů.

Diskusi obsaženou v předchozích odstavcích lze shrnout i tak, že statistické míry je sice vhodné (a při corpus-driven přístupu také nutné) použít jako základ pro hodnocení

významnosti zjištěných frekvenčních rozdílů při výběru kandidátů, tento výběr je však nakonec nutné vyhodnotit manuálně. Pro podobný postup argumentují také Rayson a Garside (2000), v popisovaném článku (Křen, 2007) se mluví o čtyřech skupinách, do nichž byla nalezená lemmata rozdělena a které byly výsledkem závěrečného vyhodnocení. Do 1. skupiny byly zařazeny odborné a jiné úzce specializované výrazy (*honitba*, *plynovod*, *souvrství*), do 2. skupiny výrazy označující témata dobové diskuse nebo odrážející technický pokrok (*euro*, *internetový*, *bosenský*), do 3. skupiny velice frekventovaná slova z centra jazyka (*on*, *se*, *můj*) a do 4. skupiny různé chyby či zkratky zvýrazněné srovnáním (*b*, *xxxx*, *foto*).

Přítomnost lemmat 4. skupiny je dána nejenom různým stupněm kvality a čištění vstupních textů (viz oddíl 4.2.5), ale také tím, že jako základ pro toto srovnání byly použity SFS, které neobsahují žádnou informaci o (ne)rovnoměrnosti rozložení jednotlivých slov v obou korpusech. Její absence je také hlavním důvodem přítomnosti odborných výrazů 1. skupiny, jejichž rozložení v korpusu je velice nerovnoměrné, a jejichž frekvence je tak rozhodujícím způsobem závislá na (ne)zařazení konkrétních textů do srovnávaných korpusů. Frekventovaná lemmata 3. skupiny byla označena většinou kvůli nenáhodné povaze jazyka, a to zejména mírami χ^2 a LL, které dávaly celkově velice podobné výsledky. Ačkoli přítomnost některých z těchto výrazů může naznačovat vývojové tendence v jazyce, je toto podezření vždy nutné ověřit. Dá se tedy říci, že lemmata zařazená do 2. skupiny jsou jediným nesporným výsledkem metody založené na SFS, jejichž hlavní nevýhodou pro diachronní srovnání je však absence informace o rozložení výskytů jednotlivých slov v korpusu.

5.2.2 Rovnoměrnost rozložení výskytů

Logickým vylepšením popsané metody je tedy přístup, který popisují Křen a Hlaváčová (2008). I v tomto případě jsou pro vyhodnocení významnosti rozdílů zjištěných mezi korpusy SYN2000 a SYN2005 použity statistické míry, tento přístup však již není založen na SFS, protože potřebuje vzít v úvahu i další informace o rozložení výskytů jednotlivých slov v korpusu.

V literatuře je popsána řada disperzních měr používaných k ohodnocení rovnoměrnosti rozložení výskytů daného slova v korpusu, jejich přehled spolu s fundovaným srovnáním jejich matematických vlastností uvádí Gries (2008). V této práci budeme ve shodě s výše zmíněným článkem (Křen a Hlaváčová, 2008) používat ***průměrnou redukovanou frekvenci*** (ARF; Savický a Hlaváčová, 2002), a to z následujících důvodů. Především se ARF prakticky osvědčila jako hlavní kritérium pro stanovování běžnosti slov při sestavování obou nejnovějších frekvenčních slovníků češtiny, *Frekvenčního slovníku češtiny* (Čermák, Křen et al., 2004) a *Frequency Dictionary of Czech* (Čermák, Křen et al., 2011). Dále je ARF v českém prostředí zažitá

díky implementaci v korpusovém manažeru Manatee/Bonito (Rychlý, 2000; Rychlý, 2007) používaném v ÚČNK a v neposlední řadě obstála ve srovnání s ostatními běžně používanými disperzními mírami (Gries, 2008).

Hodnota ARF je dána vztahem

$$ARF = \frac{1}{v} \sum_{i=1}^{f} \min\{d_i, v\}$$

kde f je frekvence daného slova v korpusu velikosti N, d_i jsou vzdálenosti mezi jednotlivými výskyty tohoto slova v korpusu a v je průměrná vzdálenost mezi výskyty daná vztahem $v = \frac{N}{f}$. Protože N je dělitelné f pouze výjimečně, ARF nabývá typicky neceločíselných hodnot, což je ale pro disperzní míry běžné. Hodnota ARF pro dané slovo je korekcí jeho frekvence založenou na rozložení jeho výskytů v korpusu: čím je rozložení rovnoměrnější, tím více se hodnota ARF blíží frekvenci a naopak; pro slova, jejichž výskyty jsou v korpusu soustředěny do jediného shluku, se hodnota ARF blíží jedné bez ohledu na frekvenci. Maximální hodnota ARF je rovna frekvenci (platí-li $d_i = v$ pro všechna i, tj. jsou-li vzdálenosti mezi všemi výskyty daného slova shodné), jeho nejmenší hodnota je rovna jedné (pro $f = 1$).

Hodnota ARF se pro velice frekventovaná funkční slova typicky pohybuje okolo poloviny jejich frekvence, ale může být i mnohonásobně (10-krát až 100-krát) menší než frekvence v případě odborných termínů vyskytujících se pouze v několika dokumentech. ARF je ve srovnání s frekvencí mnohem méně náchylná na (ne)zařazení konkrétních textů do korpusu, a lépe tedy odpovídá intuitivně chápané běžnosti slov v jazyce. Proto je také při srovnávání rozdílů mezi korpusy vhodnější používat jako základní údaj ARF než frekvenci jednotlivých slov, což ukazuje nejenom Křen (2007), ale také Baker; v jeho případě by použití ARF nebo jiné disperzní míry pravděpodobně odfiltrovalo množství odborných výrazů ve výsledcích na straně 334 (Baker, 2009a).

Pro vyhodnocení rozdílů mezi ARF byly použity tytéž statistické míry, tj. LL, χ^2 a CBF. LL doporučuje Dunning (1993) pro případy značné velikosti vzorku a relativně malých pozorovaných frekvencí, nicméně se opět potvrdilo, že jeho výsledky jsou velice podobné χ^2. Tuto podobnost, která platí zvláště pro velké vzorky, vysvětluje teoreticky Oakes (1998). Naproti tomu CBF upřednostňuje spíše méně frekventovaná slova s většími frekvenčními rozdíly oproti frekventovaným slovům s relativně malými frekvenčními rozdíly, které jsou statisticky významnější, a proto favorizované χ^2 a LL. Výsledky, které dává CBF, jsou však vhodnější pro automatické hodnocení frekvenčních rozdílů: „ ... relevancy of CBF-ranked results is less questionable. CBF can be thus suggested as a good choice for fully automatic detection, while the other measures give candidate lists suitable for further manual processing and interpretation that may come up with interesting findings.“ (Křen a Hlaváčová, 2008, str. 446)

Křen a Hlaváčová (2008) ukazují, že důsledkem použití ARF namísto frekvence je i při srovnání metodologicky shodném s předchozím (Křen, 2007) výrazné omezení vlivu odborných a jiných úzce specializovaných výrazů s velice nerovnoměrným rozložením v korpusu. Ve výsledcích proto lemmata tohoto typu skutečně chybějí, z výše zmíněných čtyř skupin tedy zůstávají kromě chyb či zkratek (4. skupina) a lemmat označujících témata dobové diskuse nebo odrážejících technický pokrok (2. skupina) hlavně velice frekventovaná lemmata z centra jazyka (3. skupina), která se pak stávají středem pozornosti článku. Zatímco přítomnost lemmat 2. a 4. skupiny je pochopitelná a u 2. skupiny i žádoucí, kladou si autoři otázku o příčinách výskytu významných frekvenčních rozdílů právě u lemmat 3. skupiny. Na srovnání frekvenčních rozdílů mezi jednotlivými hlavními typy textu (beletrie – odborná literatura – publicistika) ukazují, že zdrojem zjištěných rozdílů je ve velké většině případů publicistika. Týká se to navíc často slov typických pro beletrii, tj. takových, jejichž parciální přepočítaná ARF v beletrii je větší než v ostatních hlavních typech textu.

Autoři upozorňují, že to může být způsobeno nejenom tím, že právě publicistika je nejvíce otevřená jazykovým změnám, ale hlavně proměnami publicistiky jako žánru: rostoucí objem víkendových a jiných zájmových příloh „ředí“ původní primárně politickou orientaci a toto pronikání volnočasových témat pravděpodobně ovlivňuje jazyk publicistiky obecně, neboť se v ní tak začíná objevovat stále více neformálního jazyka typického pro beletrii (Křen a Hlaváčová, 2008, str. 445). Tytéž změny, které byly popisovány na publicistice jako celku, platí navíc i tehdy, pokud se omezíme na jediné periodikum, konkrétně Mladou frontu DNES. Toto zjištění mimo jiné ukazuje, že pozorované proměny publicistiky nejsou způsobeny subjektivitou nebo nekonzistencí anotace popisované v oddílu 4.2.4, což je důležité nejenom pro věrohodnost celého konceptu reprezentativnosti korpusů řady SYN, který je na této anotaci založen, ale samozřejmě i pro výsledky srovnávání těchto korpusů v této práci.

5.3 Vylepšení původních metod

5.3.1 Srovnávání po hlavních typech textu

Vzhledem ke zveřejnění korpusu SYN2010 je pochopitelně žádoucí aplikovat výše popisované metody na celou trojici korpusů SYN2000, SYN2005 a SYN2010. Toto rozšíření datové základny a její nové, konzistentní zpracování zahrnuté do korpusu SYN vyžaduje především úpravu způsobu vyhodnocování frekvenčních rozdílů mezi původně pouze dvěma hodnotami na hodnoty tři. Kromě toho však bylo nutné přistoupit k dalším dvěma podstatným vylepšením, jejichž popis je hlavním tématem této podkapitoly.

První vylepšení spočívá ve vyhodnocování statistické významnosti frekvenčních rozdílů odděleně po jednotlivých hlavních typech textu, ne tedy mezi korpusy jako celky, jako tomu bylo dosud. Důvodů pro tuto změnu je několik: první je teoretický a vychází z toho, že hlavní typy textu tvoří celky, které jsou homogennější než celé reprezentativní korpusy, a které má tudíž smysl zkoumat odděleně (viz zejména podkapitola 2.3). Ačkoli nejde o nijak novou myšlenku, bývají korpusy přesto srovnávány většinou jako (často poměrně heterogenní) celky. Jejich rozdělení na menší části je také přirozenější v tom, že užívané lexémy v nich mohou mít méně odlišných významů a že je také menší pravděpodobnost interference případného termínového užití v jednom hlavním typu textu s obecným užitím v jiném. Kromě toho lze očekávat, že se jazykové změny mohou týkat jen některé variety, jak zdůrazňují například Biber (1995) nebo Biber a Finegan (2001). Podobně Hundt a Mair (1999) srovnávají publicistiku (press) s odbornou literaturou (academic prose) v brownovských korpusech a docházejí k závěru, že se liší v otevřenosti přijímat jazykové změny: publicistika je progresivnější, zatímco odborná literatura konzervativnější.

Další důvod je praktický. Nelze totiž vyloučit situaci, že zatímco frekvence daného slova v jednom z hlavních typů textu roste, může ve zbývajících dvou oscilovat nebo se i mírně snižovat, což se by při srovnání frekvencí mezi korpusy jako celky nemuselo vůbec projevit. Tato situace přitom není pouze teoretická, Křen a Hlaváčová (2008, str. 443) již empiricky ověřili, že: „In most cases, the most remarkable NARF difference can be found in the newspapers, occasionally supported also by the other text registers.“ Mair et al. (2002, str. 251) k podobné situaci uvádějí: „It could be that the overall impression of stability is merely a reflection of the fact that these diachronic shifts in frequency are drowned out by much greater synchronic 'noise' generated by variation based on genre and text-type.“ Je-li však frekvenční rozdíl výrazný a je-li zaznamenán ve všech hlavních typech textu a ve všech obdobích, je pozorovaná jazyková změna tím průkaznější. Takto argumentují opět jak Křen a Hlaváčová (2008, str. 445): „Their high NARF value seems to be a meaningful relevance criterion if supported also by NARF differences observed in all the main text registers.“, tak i Leech (2004, str. 71) a Mair et al. (2002, str. 251): „The uniform increase of nouns across subcorpora shows consistency, strengthening the conviction that this is a reliable finding.“

Konečně posledním důvodem pro srovnávání frekvencí po hlavních typech textu je fakt, že jediný údaj o frekvenci slova v celém korpusu nemá velkou vypovídací hodnotu. Při srovnávání takových frekvencí totiž může výrazný rozdíl ve složení SYN2000 na jedné straně a SYN2005 a SYN2010 na straně druhé způsobovat anomálie, které lze ukázat například na lemmatu *pokud* v tabulce 5.3.1. Přestože jeho normalizovaná ARF ve všech hlavních typech textu výrazně roste, při srovnání normalizované ARF mezi celými korpusy SYN2000 a SYN2005 naopak klesá. To je způsobeno tím, že spojka *pokud* je nepříliš typická pro beletrii, její relativní ARF v ní je zhruba poloviční oproti

ostatním dvěma hlavním typům textu. Protože ale podíl beletrie mezi SYN2000 a SYN2005 vzrostl z 15 % na 40 %, došlo k výraznému poklesu ARF všech pro beletrii typických slov, který v případě spojky *pokud* nebyl kompenzován ani významným nárůstem ARF ve všech ostatních hlavních typech textu. Poznamenejme přitom, že se tomuto problému nepodařilo vyhnout ani použitím ARF, ani její normalizací. Pomohlo by pouze použití celkových přepočítaných frekvencí, které však již mění poměry hlavních typů textu v korpusech.

Podstata problému totiž spočívá v tom, že obraz stavu jazyka představovaný korpusem SYN2000 (a projevující se mj. jako poměr mezi hlavními typy textu v něm) neodpovídá obrazu stavu jazyka představovaného korpusem SYN2005. Je přitom velice nepravděpodobné, že by pozorovaný pokles ARF spojky *pokud* mezi SYN2005 a SYN2000 odpovídal tomu, že toto slovo skutečně bylo okolo roku 2005 v psaném jazyce méně běžné než okolo roku 2000. Pak ovšem nezbývá než připustit vážné pochybnosti o reprezentativnosti korpusů řady SYN jako celku, a tím i konceptu recepce a na něm stojící vyváženosti jednotlivých zdrojů. Tyto pochybnosti se ale takto zásadním způsobem nedotýkají diachronního srovnání prováděného po hlavních typech textu. Změny v podílu zařazení beletrie, odborné literatury a publicistiky jsou totiž jednoznačně nejvýraznější, a přitom je tento způsob srovnání obchází tak, že na jeho výsledky nemají žádný vliv.

	SYN2000	SYN2005	SYN2010
beletrie	14 173	16 774	24 496
odborná	31 478	36 692	45 884
publicistika	33 749	37 312	38 900
celý korpus	29 718	28 000	34 333

Tabulka 5.3.1: Normalizovaná ARF lemmatu *pokud* v hlavních typech textu.

Rozhodli jsme se proto pro srovnání subkorpusů reprezentativních korpusů odpovídající jednotlivým hlavním typům textu, tedy pro srovnání po subkorpusech repre_TTT_KKKK, nikoli repre_KKKK (viz podkapitola 4.5). Protože vycházíme ze svébytnosti všech hlavních typů textu, budeme srovnávat vždy tři trojice (repre_bel_KKKK, repre_odb_KKKK a repre_pub_KKKK) zvlášť.

V původních pracích však byly při výpočtu χ^2, LL a CBF srovnávány pouze dvě hodnoty, a to přepočítané frekvence nebo ARF v celých korpusech SYN2000 a SYN2005. Tyto hodnoty pak byly pro každé slovo vyhodnoceny trojicí měr χ^2, LL a CBF, které tak hodnotily významnost frekvenčního rozdílu; výsledné hodnoty byly použity pro určení pořadí jednotlivých slov podle jednotlivých měr ve výsledných tabulkách (Křen, 2007, str. 114–116; Křen a Hlaváčová, 2008, str. 440–442). Kontingenční tabulky byly

sestaveny shodně s článkem (Rayson a Garside, 2000), a to následujícím způsobem (a je frekvence slova w v korpusu A o velikosti N_A, b je frekvence slova w v korpusu B o velikosti N_B):

	korpus A	korpus B	součet (R_i)
w	a	b	$a+b$
non w	$N_A - a$	$N_B - b$	$(N_A + N_B) - (a+b)$
součet (S_j)	N_A	N_B	$N_A + N_B$

Tabulka 5.3.2: Kontingenční tabulka: pozorované frekvence ($O_{i,j}$).

Očekávané frekvence ($E_{i,j}$) pak byly pro každou buňku tabulky v řádku i a ve sloupci j vypočteny standardním způsobem jako

$$E_{i,j}=\frac{R_i \cdot S_j}{N_A + N_B}$$

kde R_i je součet hodnot pozorovaných frekvencí v řádku i a S_j součet hodnot pozorovaných frekvencí ve sloupci j. Platí přitom, že $E_{1,1} + E_{1,2} = O_{1,1} + O_{1,2} = a + b$. Jsou-li navíc velikosti obou srovnávaných korpusů stejné, tj. platí-li $N_A = N_B$, jsou očekávané frekvence slova w v obou korpusech stejné a rovné aritmetickému průměru pozorovaných frekvencí, tj. $E_{1,1} = E_{1,2} = \frac{a+b}{2}$.

Při rozšíření těchto kontingenčních tabulek na obecně n hodnot bylo zvažováno více možností. Oakes a Farrow (2007) sestavují při srovnávání frekvencí všech slovních tvarů v pěti různých korpusech pomocí χ^2 jednu velkou tabulku, která má v řádcích jednotlivá slova a ve sloupcích jejich pozorované frekvence ve všech srovnávaných korpusech. Očekávané frekvence pak počítají standardním, výše zmíněným způsobem na základě předpokladu, že by dané slovo mělo mít ve všech korpusech stejnou relativní frekvenci. Pokud by všechny srovnávané korpusy byly stejně velké, byla by tato očekávaná frekvence v každém z nich rovna aritmetickému průměru pozorovaných frekvencí. Použití podobně sestavené kontingenční tabulky na korpusy sledující jazykový vývoj se však neosvědčilo. Při její aplikaci na subkorpusy řady repre_bel_KKKK, repre_odb_KKKK, repre_pub_KKKK, a hlavně pub_RRRR a mf_RRRR, se totiž ukázala příliš velká preference frekventovaných slov.

Pro obecně n hodnot f_i, kde $i \in \{1, ..., n\}$, jsme proto nakonec použili ***iterativní srovnávání*** pomocí téže kontingenční tabulky o rozměrech 2×2 tak, že byl nejprve pomocí dané statistické míry (χ^2, CBF, LL) vyhodnocen rozdíl mezi f_1 a f_2, dále rozdíl mezi f_2 a f_3 atd. až po f_{n-1} a f_n. Celková hodnota pro všech $n-1$ srovnání pro danou míru je dána součtem těchto $n-1$ hodnot. Iterativní srovnávání tedy předpokládá, že frekvence očekávaná v roce $i+1$ by měla odpovídat frekvenci pozorované v roce i, frekvence očekávaná v roce $i+2$ frekvenci pozorované v roce $i+1$ atd.

Pro úplnost uvádíme přesné vzorce jednotlivých měr, podle nichž byla každá jednotlivá kontingenční tabulka vyhodnocena (kromě CBF jde o standardní míry se standardním způsobem výpočtu):

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

$$CBF = \frac{\chi^2}{\sqrt{O_{1,1}+O_{1,2}}}$$

$$LL = 2 \sum_{i,j} O_{i,j} \log \frac{O_{i,j}}{E_{i,j}}$$

Pro úplnost uveďme, že se pro výpočet χ^2 se někdy používá Yatesova korekce. Její hlavní význam je v lepší aproximaci diskrétního rozdělení pozorovaných jevů spojitým v případě jejich příliš malé frekvence. Yatesovu korekci však v této práci nepoužíváme, a to jednak proto, že není obecně přijímaným standardem (Evert, 2004, str. 82), ale hlavně proto, že podle našeho názoru není potřeba. Dodržujeme totiž podmínku, podle níž by všechny očekávané frekvence měly být větší nebo rovny 5 (Oakes, 1998, str. 25; Oakes, 2009, str. 165), takže je-li některá z očekávaných frekvencí v celém iterativním srovnání nižší než 5, je dané slovo vyřazeno a ve vyhodnocení se vůbec neobjeví. Tato podmínka přitom naše zkoumání nijak neomezuje, protože se stejně zaměřujeme na slova frekventovanější.

Celkovou hodnotu iterativního srovnání dané řady subkorpusů budeme dále pro případ použití χ^2 označovat bezpatkovým písmem jako ***chi***, pro CBF jako ***cbf*** a pro LL jako ***ll***. Bude tak označena konkrétní varianta iterativního srovnání jako metoda hodnotící průběh každého slova v dané řadě subkorpusů, ne tedy pouze jednotlivý statistický test. Zdůrazněme však, že při iterativním srovnávání subkorpusů repre_TTT_KKKK se řady repre_bel_KKKK, repre_odb_KKKK a repre_pub_KKKK vyhodnocují zvlášť jako 3 sledy 3 hodnot (beletristický, odborný a publicistický sled pro roky 2000, 2005 a 2010), takže celkový počet provedených srovnání je $3 \cdot 2 = 6$.

Iterativní srovnávání se ukázalo prakticky výhodným způsobem srovnání frekvenčního průběhu n hodnot zvýrazňujícím slova, jejichž frekvence osciluje. Ačkoli je však formulováno obecně pro libovolně velké n a bylo prakticky použito na subkorpusy pub_RRRR a mf_RRRR, tedy sledy 18 hodnot, jsou pro srovnávání delších sledů obecně vhodnější jiné metody. Přesto považujeme iterativní srovnávání za vhodný doplněk metod založených na Kendallovu τ favorizujících naopak slova s pravidelným frekvenčním nárůstem či poklesem, které budou popsány dále.

5.3.2 Srovnávání uvnitř publicistiky po jednotlivých letech

Dosud popisované vylepšení spočívající ve vyhodnocování statistické významnosti frekvenčních rozdílů odděleně po hlavních typech textu je dále doplněno vyhodnocením významnosti frekvenčních rozdílů pouze uvnitř publicistiky (případně ještě omezené na jediný titul), a to po jednotlivých letech a bez ohledu na původní zařazení těchto textů do konkrétních korpusů řady SYN. Toto druhé vylepšení vychází z již zmíněného faktu, že publicistika je jednak ze všech hlavních typů textu nejvíce otevřená jazykovým změnám, a kromě toho je u publicistiky jasně dán rok vzniku textu, který odpovídá roku vydání. Tento způsob je tedy metodologicky čistší proto, že odpadá nejenom problém s chápáním reprezentativnosti a vyváženosti, ale i se zahrnováním reedic starších textů do beletrie i odborné literatury, které rozvolňují zařazení celé kategorie v čase (viz oddíl 4.3.2). Daní za tento přístup je ovšem omezení vypovídací hodnoty jeho výsledků, které tak většinou nejsou odrazem přímo vývoje jazyka, ale často jenom změn mediálního diskursu. Přesto jsou jasně definované složení a relativně homogenní zdrojová data velkou výhodou, která již byla zmíněna v oddílu 3.3.5.

Při srovnávání frekvenčního průběhu jednotlivých slov v mnoha časových bodech však není zřejmý vhodný způsob vyhodnocení: „Since the comparison of frequency values over multiple periods of time is a relatively recent practice, there are few agreed-upon standards of how observed frequency changes in diachronic data should be statistically interpreted.“ (Hilpert a Gries, 2009, str. 385)

Zvláště připomeneme-li si tabulku 5.2.1 na straně 64 a s ní spojenou diskusi o hodnocení významnosti rozdílů mezi dvěma body na časové ose, je zřejmé, že při zobecnění na více bodů problémů jen přibude. Obtížné tak může být už jen posuzování toho, co je vývojový trend a co pouze odchylka, bez ohledu na hodnocení jeho síly či významnosti: „Trivial as this may seem, it is not always obvious whether an observed trend constitutes a significant development or an accidental fluctuation in the data.“ (Hilpert a Gries, 2009, str. 386). Také Millar (2009, str. 208) zdůrazňuje, že nelze reálně očekávat hladký a pravidelný frekvenční průběh pozorovaných jevů v čase.

Hilpert a Gries (2009, str. 389–390) doporučují použít jako míru korelace mezi dvěma sadami hodnot Kendallovo τ, které považují za vhodnější než Pearsonův korelační koeficient. Kendallovo τ totiž není tak citlivé na extrémní hodnoty a navíc jde o neparametrický test, nepředpokládá tedy žádné konkrétní rozdělení. Způsob výpočtu jsme převzali z Nelsen (2001), a to včetně korekce pro případ shodných hodnot, která je někdy označována jako tau-b.

Mějme sadu hodnot označených (x_1, y_1), (x_2, y_2), ... ,(x_n, y_n). Pro libovolný pár (x_i, y_i) a (x_j, y_j), kde $i, j \in \{1, ..., n\}$ pro $i \neq j$, potom řekneme, že je ***souhlasný*** (concordant), platí-li buď $x_i < x_j$ a zároveň $y_i < y_j$, nebo $x_i > x_j$ a zároveň $y_i > y_j$. Tento pár nazveme ***nesouhlasný*** (discordant), platí-li $x_i < x_j$ a zároveň $y_i > y_j$, nebo

$x_i > x_j$ a zároveň $y_i < y_j$. Je-li x_i=x_j nebo $y_i = y_j$, mluvíme o ***nerozhodnutém*** (tied) páru. Označíme-li c počet všech souhlasných dvojic a d počet všech dvojic nesouhlasných, je Kendallovo τ pro uvedenou sadu hodnot dáno vztahem

$$\tau(x, y) = \frac{c - d}{c + d}$$

Lze snadno nahlédnout, že $-1 \leq \tau \leq 1$. Hodnota -1 znamená největší možnou negativní korelaci mezi sadami x a y (všechny hodnoty v jedné řadě klesají, zatímco ve druhé stoupají), hodnota 1 největší možnou pozitivní korelaci (hodnoty klesají nebo stoupají v obou řadách současně), hodnota okolo 0 (téměř) žádnou korelaci. Protože celkový počet všech možných dvojic $c+d$ lze vyjádřit také jako $\binom{n}{2} = \frac{n(n-1)}{2}$, je možné vzorec upravit na

$$\tau(x, y) = \frac{2\,(c - d)}{n\,(n - 1)}$$

V případě, že se v datech vyskytují nerozhodnuté páry, je hodnota Kendallova τ modifikována takto:

$$\tau(x, y) = \frac{c - d}{\sqrt{\left(\frac{n(n-1)}{2} - T\right)\left(\frac{n(n-1)}{2} - U\right)}}$$

kde

$$T = \sum_i \frac{t_i\,(t_i - 1)}{2}$$

a

$$U = \sum_j \frac{u_j\,(u_j - 1)}{2}$$

pro t_i udávající počet nerozhodnutých párů veličiny x v i-té skupině jejích hodnot a u_j udávající počet nerozhodnutých párů veličiny y v j-té skupině jejích hodnot. Je zřejmé, že není-li ve vzorku žádný nerozhodnutý pár, a tedy $T = 0$ a $U = 0$, dostáváme opět původní vzorec. Také tato modifikace nabývá hodnot $-1 \leq \tau \leq 1$ se stejným významem.

Hilpert a Gries (2009, str. 389–390) použili Kendallovo τ na frekvenční data z korpusu TIME, analogicky je lze aplikovat i na data získaná ze subkorpusů řady pub_RRRR a mf_RRRR. Sadu $(x_1, x_2, x_3, ..., x_n)$ v tomto případě představují hodnoty $(1992, 1993, 1994, ..., 2009)$ a sadu $(y_1, y_2, y_3, ..., y_n)$ hodnoty normalizované ARF pro daný rok vydání. V tomto konkrétním případě je tedy $n = 18$ a pro $i < j$ vždy platí $x_i < x_j$.

Hlavní výhodou Kendallova τ je, že poskytuje jedinou hodnotu, která charakterizuje pravidelnost pozorované vývojové tendence. Podle této hodnoty je možné tendence vzájemně srovnávat a řadit, což je jinak bez „převedení“ grafů na čísla nemožné. Při srovnávání dvou daných průběhových grafů pouhým okem a určování větší či menší pravidelnosti pozorovaných tendencí totiž vůbec nemusíme dojít k jednoznačným výsledkům, nehledě k tomu, že je pak nemyslitelné použít corpus-driven metodu, která v korpusu najde slova s nejpravidelnější vývojovou tendencí.

Na druhou stranu je nevýhodou Kendallova τ fakt, že nebere v úvahu absolutní hodnotu zjištěných rozdílů; jinými slovy je podstatné pouze to, zda jde o průběh klesající, rostoucí či oscilující, ne už jeho síla vyjádřená v našem případě zjištěnou frekvencí. Vezměme jako příklad fiktivních frekvencí sady $(y_1, y_2, y_3, ..., y_{18}) = (1, 2, 3, ..., 18)$ a $(z_1, z_2, z_3, ..., z_{18}) = (1, 2, 4, ..., 2^{17})$ a srovnejme je pomocí Kendallova τ s $(x_1, x_2, x_3, ..., x_{18})$=(1992, 1993, 1994, ..., 2009). Protože pro $i < j$ platí nejenom $x_i < x_j$, ale i $y_i < y_j$ a $z_i < z_j$, je vždy $d = 0$ a dostáváme $\tau(x, y)$=$\tau(x, z)$=1. V obou případech tedy jde o maximální pozitivní korelaci, ačkoliv je vývojová tendence vyjádřená nárůstem frekvencí podle z zřejmě silnější než podle y. Kdybychom navíc prohodili hodnoty z_1 a z_2, bylo by $\tau(x, z) < \tau(x, y)$, i když toto prohození frekvencí 1 a 2 může být dáno pouhou náhodou. Stejné hodnoty Kendallova τ jako v předchozím případě bychom přitom dosáhli i prohozením hodnot z_{17} a z_{18}, tedy frekvencí 65 536 a 131 072, což už ale v případě reálných frekvencí pouhá náhoda být nemůže.

Hodnotu Kendallova τ jsme se proto pokusili několika způsoby korigovat tak, aby výsledné uspořádání dané touto jeho modifikovanou hodnotou bralo v úvahu i absolutní frekvenci zjištěných rozdílů, a to zapojením mediánu, rozdílu i podílu nejvyšší a nejnižší pozorované hodnoty. Přestože byl charakter výsledků v jednotlivých případech různý, nemůžeme říci, že bychom dospěli k jednoznačnému vylepšení Kendallova τ. Právě díky rozdílnému charakteru výsledků jsme se však rozhodli do závěrečného vyhodnocení zařadit kromě původní hodnoty Kendallova τ (dále označované jako ***tau***) i jednu jeho empiricky vzniklou modifikaci, a to ***taumed***. Vznikla jako vynásobení hodnoty Kendallova τ desátou odmocninou mediánu hodnot $(y_1, y_2, y_3, ..., y_n)$, jimiž jsou pro každý rok vydání normalizované ARF. Motivace je v tomto případě zřejmá, jde o přiměřené zvýhodnění frekventovaných slov s pravidelnou vývojovou tendencí.

Závěrem poznamenejme, že existují metody, které umožňují zkoumat vývojové tendence v monitorovacích korpusech ještě podrobněji a sofistikovanějším způsobem. Například lze rozdělit průběhový graf každého lemmatu na clustery odpovídající historickým obdobím, a to podle charakteristických rysů zaznamenaného frekvenčního nárůstu nebo poklesu (Hilpert a Gries, 2009, str. 390–393). Dvě lemmata s touž (nebo velice podobnou) hodnotou Kendallova τ totiž mohou mít úplně jiný průběhový graf, takže zavedení dalších charakteristik frekvenčního průběhu může pomoci při automatické kategorizaci pozorovaných vývojových tendencí. Tato podrobná charakterizace by

však už přesahovala rámec práce a pro výsledky uváděné v následující části by byla i nadbytečná, protože jejich další vyhodnocení je manuální, spojené s nahlížením do korpusu, a výsledná interpretace vychází i ze znalosti složení korpusu. Tyto možnosti však vypovídají o obtížnosti převedení celého grafu vývojové tendence na jediné číslo, které by vyjadřovalo její směr, pravidelnost, výraznost, rozdělení na fáze, tvar křivky apod.

5.3.3 Shrnutí

Z dosavadní diskuse v této kapitole vyplývá několik závěrů. Především je namísto frekvence potřeba používat ARF, kterou je kvůli srovnatelnosti žádoucí uvádět v normalizované podobě (ne však nutně jako přepočítanou frekvenci ve smyslu oddílu 5.2.1). Dále není vhodné srovnávat ARF v celých reprezentativních korpusech, ale spíše po hlavních typech textu v nich, nebo přímo její frekvenční průběh v rámci publicistiky.

Z dosavadního textu také vyplynulo celkem pět konkrétních metod (iterativní *cbf*, *chi*, *ll* a *tau*, *taumed* založené na Kendallově τ), které bychom chtěli použít na řady subkorpusů popsané v podkapitole 4.5 a výsledky vyhodnotit. Ačkoli aplikace každé z metod na každou řadu subkorpusů není problém technický, výsledné množství kombinací je příliš velké na manuální vyhodnocení, i když nebudeme uvažovat další důležité podvarianty, z nichž každá by jejich počet zdvojnásobila: vyhodnocení na slovních tvarech nebo lemmatech, s vyřazením proprií nebo bez něj. Protože by diskuse tolika variant nebyla účelná, je zřejmé, že budeme muset volit a vybrat pouze některé kombinace, které na základě dosavadní práce a empiricky provedených experimentů považujeme za nejpřínosnější.

Jak vyplývá z oddílu 5.3.1, nemá valný smysl srovnávat ARF v celých reprezentativních korpusech, takže subkorpusy řady repre_KKKK byly nakonec ze srovnání vyřazeny. Dále neuvažujeme aplikaci metod založených na Kendallově τ na subkorpusy řady repre_TTT_KKKK, protože jde v jejich případě pouze o sledy tří hodnot. Konečně jsme se na základě diskuse v oddílu 5.2.2 rozhodli seskupit výsledky všech iterativních metod aplikovaných na danou řadu subkorpusů do jedné výsledné tabulky setříděné podle *cbf*, čímž došlo de facto k volbě *cbf* jako jejich reprezentanta. V kapitole 6 tedy budeme mluvit většinou pouze o *cbf*, ačkoli výsledné tabulky iterativních metod uvádějí pro informaci také pořadí podle *chi* a *ll*.

Vzniklo tak celkem sedm kombinací uvedených v tabulce 5.3.3 a jak ukážeme dále, použité metody se vhodně doplňují, protože každá z nich nachází jiný typ výsledků: *tau* slova s pravidelným nárůstem či poklesem, *taumed* slova frekventovaná a iterativní *cbf*, *chi*, *ll* slova s velkou oscilací.

Dále bylo potřeba zvážit, zda srovnávat frekvence slovních tvarů nebo lemmat. Vyhodnocované metody dávají v obou případech velice podobné výsledky, což potvrzuje

	cbf chi ll	*tau*	*taumed*
repre_TTT_KKKK	✓		
pub_RRRR	✓	✓	✓
mf_RRRR	✓	✓	✓

Tabulka 5.3.3: Přehled subkorpusů a použitých srovnávacích metod.

i Křen (2007). Jak je vidět na výsledcích v kapitole 6, zjištěné frekvenční změny se i u ohebných slovních druhů dotýkají spíše lemmatu jako celku a nebývají omezeny jenom na jeden nebo dva jeho tvary. Protože je dostatečně spolehlivá také lemmatizace (viz oddíl 4.2.7), považujeme za vhodnější aplikovat uvedené metody na celá lemmata, aby se obraz jazykových změn při srovnávání po jednotlivých tvarech zbytečně netříštil. Použitím lemmat namísto slovních tvarů navíc odpadá otázka, zda brát ohled na velikost písmen (case-sensitivity).

Rozhodli jsme se také pro vyřazení proprií ze všech dále vyhodnocovaných srovnání, zejména vzhledem k velké proměnlivosti jejich frekvence, její závislosti na čase a zařazení konkrétních textů, a tedy také snížené výpovědní hodnoty. Propria definujeme jako lemmata začínající velkým písmenem, i když si uvědomujeme, že tato definice přináší přes svoji zdánlivou jednoznačnost závislost na konkrétní implementaci lemmatizace.

Všechna výše popsaná rozhodnutí byla aplikována jak na lexikální úrovni, tak na úrovni lexikálních kombinací, na níž však přibývají další možnosti parametrizace, a tedy i další aplikační varianty popisované v oddílu 5.4.2.

5.4 Použitá metoda

5.4.1 Úroveň lexikální

Celá dále popisovaná metoda byla implementována sadou perlovských skriptů. Tyto skripty nejdříve pro každý subkorpus z řady repre_TTT_KKKK, pub_RRRR a mf_RRRR (viz podkapitola 4.5) vypočtou ARF[1] všech lemmat, která obsahují alespoň jeden alfabetický znak a zároveň neobsahují číslici. Lemmata, která tuto podmínku nesplňují (většinou jde o čísla nebo interpunkci), jsou při výpočtech ignorována. Pro každou řadu subkorpusů repre_TTT_KKKK, pub_RRRR a mf_RRRR je výsledkem této části výpočtu tabulka o řádově stovkách tisíc řádků; každý její řádek odpovídá jednomu lemmatu vyskytujícímu se v některém ze subkorpusů dané řady, ve sloupcích je uvedena ARF tohoto lemmatu v každém konkrétním subkorpusu.

[1] Autorkou původní verze modulu pro výpočet ARF je Jaroslava Hlaváčová.

Tato tabulka je nyní statisticky vyhodnocena, a to postupně po řádcích, tj. jednotlivých lemmatech. Jde-li o tabulku pro subkorpusy řady repre_TTT_KKKK, jsou na ni aplikovány pouze *cbf*, *chi*, *ll*, a to zvlášť po jednotlivých typech textu (viz oddíl 5.3.1). Jde-li o tabulku pro subkorpusy řady pub_RRRR nebo mf_RRRR, jsou kromě *cbf*, *chi*, *ll* aplikovány také *tau* a *taumed*. Teprve v této fázi jsou vyřazena vlastní jména, v případě výpočtu *cbf*, *chi*, *ll* jsou vyřazena také lemmata, u nichž se v některé z kontingenčních tabulek iterativního srovnání objevila očekávaná ARF menší než 5 (viz opět oddíl 5.3.1)

Nakonec jsou všechna zbývající lemmata podle výsledných hodnot setříděna a v takto vzniklém pořadí také vytištěna spolu s ARF v jednotlivých subkorpusech normalizovanými na 100 milionů pozic a zaokrouhlenými na celá čísla. K této normalizaci však dochází až před tiskem, s normalizovanými hodnotami tak počítají pouze *tau* a *taumed*, zatímco statistiky *cbf*, *chi*, *ll* pracují vždy se skutečnými hodnotami ARF zjištěnými v jednotlivých subkorpusech. Protože se velikost jednotlivých subkorpusů pohybuje řádově od jednotek do stovek milionů, může se provedená normalizace ARF na 100 milionů zdát nepřiměřená. Přesnost uváděných čísel tím však není nijak dotčena, protože ARF je racionální číslo, takže i při normalizaci na 1 milion slov by *tau* a *taumed* srovnávaly totéž pořadí.

V této souvislosti je důležité poznamenat, že jsme si vědomi nevhodnosti uvádění normalizovaných nebo jinak upravených frekvencí v kontingenčních tabulkách (viz také Oakes, 2009, str. 165). Neupravené, prosté frekvence jsou však nevhodné z jiných, podstatnějších důvodů (viz oddíl 5.2.2), a proto považujeme statistické testy založené na nenormalizovaných ARF za vhodný kompromis. Protože je navíc ARF vždy nižší než frekvence (nebo je jí nanejvýš rovna), jsou pozorované i očekávané frekvence jednotlivých slov vyšší než odpovídající ARF (samozřejmě za předpokladu, že „očekávaná rovnoměrnost rozložení výskytů“ bude odpovídat rovnoměrnosti pozorované), takže reálně pracujeme s čísly i v tomto smyslu spolehlivějšími. Kromě toho s odvoláním na diskusi v oddílu 5.2.1 znovu konstatujeme, že naším cílem není přesné určení statistické významnosti zjištěných frekvenčních rozdílů, ale především nalezení a vyhodnocení vhodných způsobů jejich uspořádání.

V neposlední řadě nám nic jiného než práce přímo s ARF nezbývá, ačkoli by samozřejmě bylo možné pracovat s frekvencí namísto ARF s tím, že bychom slova rozložená v korpusu „nedostatečně rovnoměrně“ na závěr pouze odstranili. Tato nerovnoměrnost by mohla být dána prahovou hodnotou Juillandova D nebo jiného podobného koeficientu, případně kombinovaného s minimálním počtem dokumentů v korpusu, v nichž by se dané slovo muselo vyskytovat. Tento přístup používají Oakes a Farrow (2007, str. 93), prahovou hodnotu Juillandova D stanovili empiricky na 0,3; znamená to však, že zatímco slova s hodnotou jen o málo vyšší (např. 0,31) nakonec ze závěrečných tabulek odstraněna nejsou, slova s hodnotou naopak nižší (např. 0,29) odstraněna jsou,

ačkoli míra nerovnoměrnosti rozložení výskytů je v obou případech přibližně stejná. Snažíme se proto používání prahových hodnot vyhnout, je-li to jen trochu možné, protože se domníváme, že graduální znevýhodňování frekvencí nerovnoměrně rozložených slov pomocí ARF je pro tyto účely vhodnější a převáží jeho případné jiné nevýhody.

5.4.2 Úroveň lexikálních kombinací

Na této úrovni se budeme snažit najít konkrétní lexikální kombinace, jejichž frekvence ve sledovaných subkorpusech vykazují největší změny. Výsledky budou vyhodnoceny a prodiskutovány společně s výsledky získanými na lexikální úrovni. Pod pojmem lexikální kombinace rozumíme v zásadě kolokace; protože však jde o velmi rozsáhlé téma ležící mimo oblast zájmu této práce, nechceme se na tomto místě pouštět do podrobných rozborů týkajících se vymezení pojmu kolokace a různých způsobů jejich vyhledávání. Problémem je už nejasná definice samotného pojmu, velká různorodost kolokací jak z lingvistického (Čermák, 2001b), tak matematického hlediska (Evert, 2004), a z toho plynoucí praktické problémy při jejich identifikaci.

Tato různorodost se projevuje mimo jiné tím, že jednotlivé statistické míry prakticky používané pro jejich vyhledávání mohou preferovat jiný typ kolokací, takže není výjimkou ani situace, kdy se výsledky jednotlivých měr diametrálně liší. Různý charakter výsledných kolokátů získaných při aplikaci odlišných statistických měr na češtinu popisuje Křen (2006b), který také potvrzuje známý rozdíl mezi T-score a MI-score: zatímco MI-score nachází silné kolokace s velkou relativní frekvencí, a tedy spíše výjimečné až náhodné, T-score naopak kolokace nenáhodné, pravidelné a ustálené. Protože tedy nemůžeme očekávat uspokojivé podchycení celé široce definované množiny kolokací pomocí jediné univerzální statistické míry, používají se v praxi často jejich kombinace (Pecina, 2009).

Cílem této práce však není hledání metod vhodných pro vyhledávání kolokací, ale „pouze“ snaha zachytit vývojové tendence v jazyce i na této úrovni, a to na základě korpusů zachycujících několik blízkých časových období. Ideální by přitom bylo najít takový způsob srovnávání jednotlivých korpusů, který se v maximální možné míře vyhýbá nutnosti prakticky rozlišovat, co kolokací je a co jí už není.

Proto byl implementován postup analogický srovnávání na lexikální úrovni, který se od toho původního popisovaného v oddílu 5.4.1 liší pouze tím, že namísto ARF jednotlivých lemmat bere jako základ pro srovnání ARF jejich dvojic. Hlavní výhodou tohoto postupu je, že bere v úvahu všechny dvojice slov, které se v dané řadě subkorpusů vyskytují (a frekvenční rozdíly mezi nimi), tedy nejenom dvojice považované za kolokace, a není proto nijak závislý na jejich hodnocení jednotlivými mírami. Odpadá tak problém s rozlišováním kolokací, řešením vztahu mezi jejich frekventovaností, typičností a ustáleností, volbou konkrétní statistické míry či jejich kombinace i různých pra-

hových hodnot (např. pro MI-score je zásadní volba minimální frekvence celé kolokace). Je tak sice nutné zpracovat mnohem větší množství dat, to ale technicky není problém. Další nezanedbatelnou výhodou tohoto postupu je, že umožňuje srovnávání přímo normalizovaných ARF jednotlivých dvojic, tedy údajů statisticky průkaznějších, než je pouhé pořadí kolokací v rámci jednotlivých subkorpusů, které je navíc dané jejich ohodnocením jednotlivými mírami.

Tento postup skutečně najde všechny dvojice, jejichž frekvenční rozdíly jsou v dané řadě subkorpusů signifikantní. Ukázalo se však, že velká většina těchto dvojic jsou kombinace náhodné povahy, které jistě nejsou kolokacemi ani v hodně širokém slova smyslu. Dosud popsaný postup proto musel být upraven přidáním dodatečného kolokačního filtru, který ponechá ve výběru pouze dvojice splňující následující tři kritéria: hodnota MI-score pro danou dvojici je větší nebo rovna 4, hodnota LL je větší nebo rovna 10 a zároveň je ARF této dvojice větší nebo rovna 10. Tyto prahové hodnoty jsou počítány vždy v dané řadě subkorpusů jako celku (tedy např. najednou ve všech subkorpusech řady mf_RRRR), která je pro kolokační filtr považována za jeden korpus. Předchází se tak nejasnostem, který ze subkorpusů dané řady považovat za rozhodující v případě, že by daná dvojice splňovala požadovaná kritéria jenom v některém z nich.

Popsaný kolokační filtr a jeho prahové hodnoty vycházejí ze způsobu označování kolokací použitého ve *Slovníku Karla Čapka* (Čermák et al., 2007), z něhož v tomto smyslu vychází i *Slovník Bohumila Hrabala* (Čermák, Cvrček et al., 2009). Kolokace byly ve zdrojovém korpusu Karla Čapka vyhledávány jako bigramy a označovány pomocí kombinace tří měr, a to MI-score, LL a ϕ^2 koeficientu,[2] k jejichž výpočtu bylo použito NSP (Banerjee a Pedersen, 2003). V následujícím kroku byly odstraněny bigramy, pro něž byla hodnota MI-score < 4 nebo hodnota LL < 10 nebo jejichž frekvence v korpusu byla < 3. Tyto prahové hodnoty byly stanoveny empiricky, cílem jejich použití bylo odfiltrovat náhodné slovní kombinace (pomocí prahové hodnoty pro LL) a kombinace frekventovaných slov, často spíše gramatického charakteru (pomocí prahové hodnoty pro MI-score). Každý bigram ve výsledné množině byl ohodnocen výše zmíněnými mírami, toto ohodnocení bylo převedeno na pořadí (rank) podle dané míry v celé množině a tato tři pořadí byla pro každý bigram sečtena. Všechny bigramy byly nakonec podle těchto sumárních ranků setříděny, čímž bylo dáno pořadí podle celkové významnosti kolokace odrážející často protichůdná hodnocení podle jednotlivých měr tak, aby byly ve výsledku zastoupeny kolokace různého typu. Celý postup popisuje také Křen (2008), ve srovnání s touto prací je navíc zajímavé, že v obou autorských slovnících byla významnost takto nalezených kolokací ve srovnání s jejich frekvencemi v reprezentativním korpusu SYN2005 označena pouze pomocí χ^2.

[2]Protože platí $\phi^2 = \frac{\chi^2}{N}$ a N je pro daný korpus konstanta, je uspořádání výsledků podle obou měr stejné a liší se pouze číselnou hodnotou; namísto ϕ^2 je tedy možné používat i χ^2 (viz také Křen, 2006b).

Zdůrazněme však, že jsme pro kolokační filtr převzali pouze empiricky osvědčené prahové hodnoty pro MI-score a LL, ne tedy hodnocení kolokací sumárními ranky. Požadavek na minimální frekvenci celé dvojice byl vzhledem ke korpusům výrazně větším, než jsou zmíněné autorské korpusy, přiměřeně zvýšen. Dalším odlišným rysem aplikovaného kolokačního filtru je, že namísto frekvence bere v úvahu nenormalizovanou ARF, což pomáhá odfiltrovat dvojice slov s nerovnoměrným rozložením výskytů, tedy především víceslovné odborné výrazy.

Kromě toho, že všechny dvojice musely projít výše popsaným kolokačním filtrem, je metoda použitá na úrovni lexikálních kombinací shodná s metodou použitou na lexikální úrovni a popsanou v oddílu 5.4.1, a to včetně aplikace stejných kombinací metod a subkorpusů daných tabulkou 5.3.3 na straně 77. To, že tyto metody pracují s ARF celých dvojic namísto ARF jednotlivých lemmat, je z hlediska výpočtu ARF a jejího srovnávání mezi jednotlivými subkorpusy dané řady rozdíl pouze technický.

Podstatný je však způsob výběru těchto dvojic, protože možností je celá řada. Dvojice byly vybrány jako bigramy lemmat, tedy s fixním pořadím, avšak jdoucí přes případnou hranici danou interpunkcí (ve výsledcích tedy najdeme např. kombinaci *myslet že*). Toto poměrně striktní vymezení zvyšuje precision, tj. pravděpodobnost, že výsledné dvojice jsou skutečně relevantní. Ve srovnání s dvojicemi s větší velikostí okna (a tedy i s případnými mezilehlými slovy) a proměnlivým pořadím ho považujeme za vhodnější, protože rozvolnění bigramů přináší sice další dvojice, jde však spíše než o lexikální kombinace o slova spolu volně související. Tuto volbu potvrzuje i způsob zvolený pro vyhledávání kolokací v obou autorských slovnících: také jejich detekce v korpusu byla založena pouze na bigramech lemmat (Křen, 2008, str. 477–478), i když byly před zařazením do slovníku převedeny do správného tvaru a případně rozšířeny (např. *pýcha předcházet* na *pýcha předchází pád*).

Stejně jako na lexikální úrovni platí, že všechna lemmata v bigramech obsahují alespoň jeden alfabetický znak a zároveň neobsahují číslici. Dále jsou vyřazeny bigramy obsahující alespoň jedno proprium, při výpočtu *cbf*, *chi*, *ll* navíc také bigramy, u nichž se v některé z kontingenčních tabulek iterativního srovnání objevila očekávaná ARF menší než 5 (viz oddíl 5.3.1). Nakonec jsou všechny zbývající lexikální kombinace podle výsledných hodnot dané statistiky setříděny a v takto vzniklém pořadí také vytištěny spolu s ARF v jednotlivých subkorpusech normalizovanými na 100 milionů pozic a zaokrouhlenými na celá čísla. Stejně jako na lexikální úrovni dochází k této normalizaci až před tiskem, s normalizovanými hodnotami tak počítají pouze *tau* a *taumed*, zatímco statistiky *cbf*, *chi*, *ll* pracují vždy se skutečnými hodnotami ARF zjištěnými v jednotlivých subkorpusech.

6 Výsledky a diskuse

6.1 Úvod

Kapitola s výsledky je členěna do čtyř hlavních částí: zvlášť stojí výsledky založené na reprezentativních subkorpusech, zatímco výsledky založené na publicistických subkorpusech pub_RRRR a mf_RRRR jsou probírány společně; rozdělení do jednotlivých částí je v tomto případě určeno konkrétní metodou použitou při vyhodnocování frekvenčních rozdílů. Lexikální úroveň i úroveň lexikálních kombinací jsou přitom vždy spojeny do jedné části, jednotícím prvkem je tedy použitá metoda. Toto uspořádání považujeme za nejvhodnější vzhledem k tomu, že výsledky na pub_RRRR a mf_RRRR dané stejnou metodou se často překrývají nebo jsou podobného typu. Toto překrývání navíc znamená, že lemmata jsou kombinacemi dále rozvíjena, takže úroveň lexikálních kombinací často poskytuje příklady a osvětluje výsledky zjištěné na lexikální úrovni.

Každá ze čtyř hlavních částí (podkapitol) začíná tabulkami s výsledky dané metody aplikované na zdrojové subkorpusy a pokračuje jejich diskusí, která je zakončena průběhovými grafy vhodně vybraných lemmat a kombinací ilustrujících popisované jevy.

Úvodní tabulky obsahují pro každou metodu a řadu subkorpusů celkem 50 lemmat a 50 kombinací, jejichž frekvenční průběh je podle použité metody považován za nejvýraznější. Přestože každá metoda vyhodnocuje jejich průběh jiným způsobem a lišit se může také použitá řada subkorpusů, mohou se lemmata a kombinace uváděné ve výsledných tabulkách pro danou metodu a řadu subkorpusů částečně překrývat s jinými. Každá tabulka je z prostorových důvodů rozdělena na dvě poloviny, na každé straně tedy najdeme 25 položek. V prvním sloupci každé tabulky je uveden rank (pořadí) podle dané metody (případně metod, jde-li o *cbf*, *chi*, *ll*), dále konkrétní ***výraz***,[1] v dalším sloupci znak + nebo − označující nárůst nebo pokles (pouze u *tau* a *taumed*) a nakonec řada hodnot ARF daného výrazu v konkrétním subkorpusu normalizovaných na 100 milionů pozic.

Poté následuje text s diskusí a vyhodnocením výsledků podaných v tabulkách. Jsou-li výsledky založeny na subkorpusech řady pub_RRRR a mf_RRRR, je v textu dodržována konvence v označování jednotlivých výrazů pomocí indexů M, P nebo MP podle toho, ze které tabulky daný výraz pochází. Jde tedy zároveň o jednoduchý způ-

[1] V dalším textu budeme takto souhrnně označovat lemmata a kombinace.

sob odkazování od každého výrazu v diskusi na příslušnou tabulku. Například notace $^{\mathrm{M}}$*chřipka* označuje lemma „chřipka“, které se vyskytuje mezi prvními 50 lemmaty pouze tehdy, jsou-li výsledky založeny na datech mf_RRRR (zatímco na datech pub_RRRR mezi prvními 50 není), $^{\mathrm{MP}}$*výsledek volba* označuje kombinaci, která se vyskytuje mezi prvními 50 lemmaty v obou řadách subkorpusů, a můžeme ji tedy v příslušné podkapitole najít ve dvou tabulkách. Kombinace uvádíme záměrně v základním tvaru, aby bylo zřejmé, že mohou někdy připouštět více interpretací („výsledek volby“ nebo „výsledky voleb“) a že výstup může být ovlivněn použitou lemmatizací (samostatné lemma „volby“ v datech nenajdeme, vše je lemmatizováno singulárem, na rozdíl od dvojice „novina“ a „noviny“).

Lemmata a kombinace mohou být součástí víceslovné jednotky, což však použitá lemmatizace nijak neodráží. V některých případech jde navíc o části víceslovných proprií, která – pokud by byla skutečně považována za jednotku a lemmatizována s velkým počátečním písmenem – by byla jako propria v průběhu zpracování vyřazena. Upozorněme také, že součástí lemmatu nikdy není reflexivum, které lemmatizace i v případě reflexiv tantum považuje za samostatné zájmeno.

Hlavním cílem diskuse je charakterizace metod a jejich výsledků ve vztahu ke změnám v jazyce a složení zdrojových subkorpusů; nemůže se proto podrobně věnovat každému zjištěnému výrazu. To platí zejména o podkapitole 6.2, která se jako jediná zabývá výsledky metody aplikované na reprezentativní subkorpusy. Průběh diskuse je v tomto případě netypický zaměřením nikoli na jednotlivé výrazy ve výsledných tabulkách, ale na složení jednotlivých reprezentativních subkorpusů, vztah mezi různými hlavními typy textu v nich a jejich vliv na výsledky. Pro podkapitoly 6.3, 6.4 a 6.5 založené na publicistických subkorpusech však platí, že v diskusi je zmíněn každý výraz z výsledných tabulek.

Zdůrazněme však, že diskuse není založena pouze na předkládaných tabulkách. Pro každý výraz v každé tabulce byly nejprve pro snadnější orientaci vytvořeny přehledné průběhové grafy, poté následovaly cílené sondy do konkrétních subkorpusů pomocí Bonita spojené s vyhodnocením typických užití jednotlivých výrazů a zkoumáním zdrojových textů (zvláště v místech se silnými výkyvy). Výsledkem je kategorizace výrazů, která je pro každou podkapitolu různá a která napomáhá vysvětlení příčin pozorovaných tendencí či výkyvů.

Na závěr každé části je otištěna část těchto průběhových grafů, které zobrazují frekvenční průběh daného výrazu v čase na všech subkorpusech řady pub_RRRR a mf_RRRR, tedy v letech 1992–2009. Tyto publicistické subkorpusy jsou zvoleny i tehdy, když jde o metodu aplikovanou na reprezentativní subkorpusy. Publicistika totiž nejrychleji odráží změny úzu, rok vydání v ní odpovídá roku vzniku textu a také je v ní k dispozici nejvíce dat. Využitím většího množství časových bodů na relativně velkých

datech se také snáze vyhneme nebezpečí oscilace některých jevů a z ní pramenícím nespolehlivým závěrům, které popisuje Millar (2009, str. 208).

V záhlaví každého grafu je vždy uveden CQL dotaz, jímž byl daný výraz vyhledán a jemuž tedy graf odpovídá. Grafy se numericky plně shodují s příslušnými tabulkami, jsou založeny na stejných datech a také výsledné frekvence jsou udávány jako ARF normalizovaná na 100 milionů pozic. Jde tedy pouze o doplňkový, přehlednější, ale také prostorově náročnější způsob zobrazení dat a prezentace výsledků.

Každý jednotlivý graf vznikl tak, že daný CQL dotaz byl přes Manatee postupně aplikován na každý ze subkorpusů, jeho výsledná ARF v něm byla normalizována, zaznamenána do tabulky a tato tabulka pak byla vynesena do grafu sadou příkazů utility `gnuplot`. Celý tento postup byl automatizován v jediném perlovském skriptu, pro dotazy na subkorpusy byla použita upravená verze knihovny Perlmanatee.pm[2] zajišťující komunikaci přímo s `manateesrv`, což je výkonná část Manatee realizující vyhledávání v korpusech. Tento přístup umožňuje automatické vyhodnocení libovolného dotazu odpovídajícího Manatee CQL a vynesení jeho frekvenčního průběhu do grafu, a to na obecně libovolné subkorpusy, ačkoli byl v tomto případě omezen jenom na subkorpusy řady pub_RRRR a mf_RRRR.

6.2 Iterativní *cbf*, *chi*, *ll* na reprezentativních subkorpusech

Tato podkapitola se zabývá rozborem výsledků iterativních metod *cbf*, *chi*, *ll* (viz oddíl 5.3.1) aplikovaných na reprezentativní subkorpusy řady repre_TTT_KKKK. Tyto subkorpusy jsou v záhlaví výsledných tabulek uváděny z prostorových důvodů pouze jako TTT_KKKK, například tedy bel_2010. Samy výsledky uváděné v tabulkách jsou velice různorodé, najdeme tu řadu výrazů velice frekventovaných (*ale*, *už*, *ten* atd.), dále výrazy, jejichž frekvenční posuny odrážejí měnící se realitu (*mobil*, *internet*, *celebrita* atd.), a také zkratky či chyby způsobené nedostatečným čištěním textů (*tel*, *stol*, *í* atd.).

Mnohými z těchto výrazů se budeme podrobněji zabývat v následujících podkapitolách. Jak však již bylo zmíněno v úvodu, diskuse v této podkapitole je poněkud netypická. Jejím cílem není podrobný rozbor výrazů ve výsledných tabulkách a jejich frekvenčního průběhu, zaměřena je spíše na složení jednotlivých reprezentativních subkorpusů a jeho vliv na výsledky. Jak se totiž ukáže, je tento vliv zásadní, což spolu s dalšími metodologickými problémy srovnání tohoto typu do značné míry zpochybňuje; rozbor jeho konkrétních výsledků pak ovšem ztrácí opodstatnění.

[2] Autorem původní verze knihovny je Václav Cvrček.

cbf	*chi*	*ll*	lemma	bel_2000	bel_2005	bel_2010	odb_2000	odb_2005	odb_2010	pub_2000	pub_2005	pub_2010
1	42	47	mobil	15	118	589	218	184	679	331	2510	2109
2	9	9	kraj	4325	4989	4130	1907	2873	3145	3128	10627	20979
3	58	59	hejtman	107	82	48	64	146	194	123	1893	3264
4	6	6	já	386663	346131	339811	24024	15978	36507	46802	69232	68714
5	24	23	tenhle	40932	45604	48359	2330	1354	6013	5392	10516	11723
6	33	37	internet	10	80	249	2934	2162	3231	1556	5700	6010
7	10	10	vy	71054	66139	68688	14623	10324	27219	15496	22303	24619
8	39	32	r	810	1068	380	10068	8247	3631	4250	1281	1213
9	78	58	foto	56	96	88	817	205	584	4117	939	1624
10	29	29	krajský	234	168	155	295	600	802	2127	6061	11063
11	234	247	bin	71	26	44	36	26	65	91	815	345
12	3	3	ale	322522	313627	327377	137260	135157	179281	166216	249791	269429
13	15	15	kvůli	9307	11195	13113	3147	3891	7624	12598	25266	29464
14	167	158	fax	36	102	129	2066	336	301	1506	609	216
15	150	141	tel	20	64	64	2471	322	347	1650	736	713
16	276	307	celebrita	31	46	147	41	85	264	142	899	1013
17	82	88	mobilní	66	154	296	995	1217	1789	1150	3918	3221
18	89	84	sezona	66	148	358	213	336	996	2481	2157	6325
19	7	7	už	179524	175809	169223	38398	37603	59176	86368	126609	141694
20	2	2	ten	1240187	1158909	1143638	419991	409211	522912	475307	615963	645187
21	23	24	teď	62317	61438	64138	4709	3876	8679	12116	21093	25035
22	189	166	ing	122	82	64	1862	456	323	1644	427	274
23	61	62	jenže	8889	8042	10075	1331	1088	3197	4438	8383	10294
24	46	45	určitě	8981	11287	14611	3188	2358	5850	6716	10627	13187
25	60	68	médium	188	264	442	2293	2273	3277	2885	7326	6756

Tabulka 6.2.1: Úroveň lexikální, *repre_TTT_KKKK*, *cbf chi ll*, 1. část.

cbf	*chi*	*ll*	lemma	bel_2000	bel_2005	bel_2010	odb_2000	odb_2005	odb_2010	pub_2000	pub_2005	pub_2010
26	136	152	mediální	25	80	127	341	538	765	895	2924	2302
27	145	147	fanoušek	148	136	211	195	176	947	2228	4403	6132
28	43	46	tady	36011	35057	39250	4310	2589	5930	8697	14228	15005
29	606	592	agresívní	107	72	18	145	88	25	228	593	117
30	52	51	jestli	23708	29668	34318	2489	2059	5103	6553	10548	11611
31	113	120	bavit	4845	4511	5374	1185	804	2195	2472	5099	5759
32	173	177	no	17473	15480	16744	1099	600	1891	1410	2908	2936
33	746	773	cédéčko	0	56	119	95	38	129	123	530	386
34	105	110	taky	28599	26133	26229	1263	1024	2859	3033	5882	6210
35	298	339	operátor	51	64	117	441	521	784	317	1305	1152
36	317	282	p	759	228	225	1939	1164	1236	1308	380	640
37	643	476	stol	61	452	34	254	225	209	100	66	86
38	178	157	marka	336	326	193	318	275	258	3489	1426	419
39	5	5	on	616752	647935	680572	158650	166277	188710	169361	238642	233947
40	153	169	hasič	408	332	310	345	152	264	1391	3614	4698
41	441	448	fotka	841	750	1134	136	97	483	294	838	1251
42	163	167	vyrazit	4356	5581	6729	622	459	1337	1642	3461	4299
43	247	265	úžasný	1518	1694	2511	472	489	1181	486	1429	1749
44	1	1	se	1923316	1996759	2086670	1022489	1057401	1199825	1064477	1298511	1308566
45	788	778	přehrávač	36	74	133	173	70	458	153	440	536
46	1292	1010	í	138	36	42	186	67	141	215	11	76
47	519	556	portál	301	212	177	223	430	861	153	633	967
48	70	70	opravdu	16291	16462	18483	5223	4116	8851	7328	11394	11865
49	225	232	takhle	9832	10705	10561	731	410	1386	1797	3427	3637
50	271	220	t	571	290	275	3016	1837	1294	1192	192	246

Tabulka 6.2.2: Úroveň lexikální, *repre_TTT_KKKK*, *cbf chi ll*, 2. část.

cbf	*chi*	*ll*	kombinace	bel_2000	bel_2005	bel_2010	odb_2000	odb_2005	odb_2010	pub_2000	pub_2005	pub_2010
1	1	1	mobilní telefon	20	74	199	286	407	682	737	2715	1916
2	3	3	devadesátý rok	76	110	111	404	737	753	1095	2810	2089
3	23	25	krajský město	20	38	20	23	59	61	75	306	759
4	2	2	tisíc koruna	143	98	80	182	243	569	6323	10904	11403
5	7	7	sdělovací prostředek	173	112	139	540	418	298	1713	583	299
6	22	17	národní rada	56	32	22	123	219	83	661	137	140
7	12	10	cenný papír	102	84	62	204	585	320	1964	1052	640
8	6	6	akciový společnost	51	58	30	727	471	271	2501	1055	759
9	120	145	rudý právo	209	42	60	100	32	46	217	124	99
10	19	18	tiskový konference	117	62	143	377	152	243	2669	1545	1551
11	43	48	kroutit hlava	306	310	352	50	23	108	314	730	607
12	214	220	šestý smysl	61	76	111	27	26	49	13	82	28
13	24	29	pár minuta	678	988	1395	132	149	286	261	656	642
14	17	19	k vidění	219	242	304	295	155	344	975	1716	2459
15	72	75	cestovní kancelář	127	148	149	490	114	249	966	941	896
16	35	38	za volant	357	404	525	82	59	191	417	904	1114
17	25	30	rodinný dům	46	50	76	127	246	400	375	825	1302
18	4	4	deset rok	1915	1786	2009	1199	1837	1688	3714	6427	5403
19	102	69	t r	36	26	12	204	129	43	190	21	20
20	115	130	přístrojový deska	46	76	88	32	41	46	47	185	81
21	154	156	n m	15	38	105	173	85	215	47	111	117
22	11	15	rok vězení	51	56	58	50	47	58	727	1716	1881
23	74	83	kreditní karta	41	62	153	82	135	157	125	348	305
24	40	42	vědět jestli	2002	2255	2493	241	129	381	677	1157	1117
25	110	112	n l	46	144	62	318	316	424	67	190	135

Tabulka 6.2.3: Úroveň lexikálních kombinací, *repre_TTT_KKKK*, *cbf chi ll*, 1. část.

cbf	*chi*	*ll*	kombinace	bel_2000	bel_2005	bel_2010	odb_2000	odb_2005	odb_2010	pub_2000	pub_2005	pub_2010
26	20	22	pár den	1324	1566	1882	236	202	473	858	1598	1774
27	67	67	dětský hřiště	46	48	50	50	47	89	89	208	571
28	116	102	provdat za	219	268	328	64	269	101	78	134	91
29	51	57	obchodní centrum	31	36	84	100	129	184	245	572	868
30	164	179	strašně moc	132	152	167	50	18	74	94	240	294
31	151	150	pan profesor	438	296	181	132	47	135	149	195	104
32	65	54	ministr zahraniční	36	34	44	32	111	55	560	198	129
33	8	8	letošní rok	97	64	88	1372	679	910	6497	4764	5921
34	352	375	válečný tažení	36	26	30	41	47	40	13	63	15
35	71	73	lidový noviny	82	74	58	204	64	154	1399	989	680
36	232	213	t j	51	16	30	354	228	89	72	24	25
37	125	146	taneční hudba	25	30	34	41	20	18	88	266	155
38	94	97	cestovní ruch	15	46	22	395	158	237	639	794	1162
39	243	241	m n	15	38	105	191	85	175	39	66	89
40	5	5	milión koruna	36	32	40	391	445	864	10646	14774	15972
41	112	122	hlavní hrdina	87	68	72	118	64	203	239	459	543
42	212	231	motorový pila	46	38	64	50	23	49	61	190	183
43	9	9	poté co	1228	2067	2907	1490	2235	2810	6906	7656	6342
44	70	68	americký film	82	44	64	50	61	92	805	482	195
45	558	615	demokratický republika	71	12	38	32	44	37	105	90	91
46	26	26	ministr zahraničí	173	88	131	41	108	95	3042	1996	1426
47	475	481	mluvit česky	97	54	46	59	12	49	85	129	102
48	42	40	generální tajemník	87	48	58	73	135	114	1180	577	340
49	46	41	zahraniční obchod	36	26	32	218	313	175	941	411	274
50	105	106	jízdní řád	173	178	139	254	94	138	329	330	685

Tabulka 6.2.4: Úroveň lexikálních kombinací, *repre_TTT_KKKK*, *cbf chi ll*, 2. část.

Ve výsledných tabulkách se hojně vyskytuje oscilace konkrétních výrazů v různých obdobích a hlavních typech textu, a proto jsme jako východisko pro další diskusi zvolili charakterizaci založenou na zjednodušeném vyjádření frekvenčního průběhu každého výrazu v rámci hlavních typů textu. Toto vyjádření je binární, hodnocené pouze jako nárůst či pokles normalizované ARF symbolizovaný šipkami, a to vždy mezi dvěma časově následujícími subkorpusy v rámci stejného hlavního typu textu. Například pro kombinaci *strašně moc*, která je v tabulce 6.2.4 na předchozí straně na řádku s *cbf* rankem 30, tak dostáváme v beletrii ↑ ↑ (nárůst mezi repre_bel_2000 a repre_bel_2005 a nárůst mezi repre_bel_2005 a repre_bel_2010), v odborné literatuře ↓ ↑ (pokles následovaný nárůstem) a v publicistice opět dvojí nárůst ↑ ↑; celou šestici tedy můžeme v pořadí beletrie, odborná literatura a publicistika zapsat jako ↑ ↑ ↓ ↑ ↑ ↑. Poznamenejme, že každá šipka zároveň odpovídá jednomu srovnání prováděnému při iterativním vyhodnocení, při němž se ovšem bere v úvahu také absolutní hodnota zjištěných frekvencí (viz opět oddíl 5.3.1).

Frekvenční průběh každého z 50 lemmat a 50 kombinací ve výsledných tabulkách byl automaticky vyhodnocen a zaznamenán touto šesticí indikátorů. Byla zjištěna především jejich velká různost, z celkem možných $2^6 = 64$ kombinací indikátorů jich bylo mezi 100 výrazy nalezeno 36 (nepočítáme-li lemma *tel* s jediným případem shody dvou následujících frekvencí). Tato různost je pravděpodobně podpořena použitou metodou, která za nejvýznamnější označuje výrazy s velkými frekvenčními rozdíly.

Tabulka 6.2.5 ukazuje všechny šestice, které byly nalezeny alespoň čtyřikrát, spolu s počty výrazů, u nichž byl daný průběh zaznamenán. Tato tabulka bude výchozím bodem další diskuse, v níž se zaměříme zejména na odhalování příčin těchto průběhů, mezi nimiž je často nevhodné složení výchozích dat.

počet	beletrie	odborná	publicistika
15	↑ ↑	↓ ↑	↑ ↑
11	↓ ↑	↓ ↑	↑ ↑
10	↑ ↑	↑ ↑	↑ ↓
6	↑ ↑	↑ ↑	↑ ↑
6	↑ ↑	↓ ↑	↑ ↓
4	↓ ↑	↑ ↓	↓ ↓
4	↓ ↓	↑ ↑	↑ ↑

Tabulka 6.2.5: Nejčastější šestice indikátorů nárůstu a poklesu.

Nejčastější je již zmíněná šestice ↑ ↑ ↓ ↑ ↑ ↑, kterou se budeme podrobněji zabývat dále a která poukazuje na rozpor mezi obecným trendem a frekvenčním rozdílem mezi subkorpusy repre_odb_2000 a repre_odb_2005. Podobného typu je také šestice ↑ ↑

↑ ↑ ↑ ↓, kde je však pokles neodpovídající ostatním subkorpusům zaznamenán mezi subkorpusy repre_pub_2005 a repre_pub_2010, a to u těchto 10 výrazů: *devadesátý rok*, *kreditní karta*, *mediální*, *mobilní*, *mobilní telefon*, *on*, *operátor*, *pár minuta*, *poté co*, *přístrojový deska*. Příčin tohoto jevu je více, jednou z nich je pravděpodobně proměnlivé složení publicistiky v reprezentativních subkorpusech, které je vidět na obr. 4.5.2 na straně 55. Upozorněme však, že problematické složení publicistických subkorpusů popisované v dalších podkapitolách se týká pouze subkorpusů řady pub_RRRR (a v menší míře také mf_RRRR).

Frekvenční průběh kombinace *pár minuta* v publicistických subkorpusech řady pub_RRRR a mf_RRRR je vidět na obr. 6.2.1 na straně 96. Protože v nich vykazuje poměrně stabilní nárůst, je otázka, čím je způsoben rozpor s poklesem zaznamenaným mezi subkorpusy repre_pub_2005 a repre_pub_2010. Kromě rozdílného složení jednotlivých subkorpusů může jít i o celkovou nespolehlivost údajů vyvozovaných na základě malého množství časových bodů. Millar (2009, str. 208) uvádí konkrétní příklady, kdy vývojová tendence vyvozená z pouhých dvou časových bodů popírala zřejmý obecný trend pozorovatelný na týchž datech, a dodává: „Changes in the frequency of language features over time clearly cannot be assumed a priori to be smooth. ... So data from multiple chronological points appear to be essential to obtain a clear overview of any trend and to allow for accurate statistical modeling.“

Za zmínku stojí také průběh ↓ ↓ ↑ ↑ ↑ ↑, který znamenává pokles v beletrii a nárůst v ostatních slovních druzích a který vykazují výrazy *hejtman*, *krajský*, *portál* a *tisíc koruna*. V některých případech se tento průběh nepodařilo uspokojivě vysvětlit (*tisíc koruna*), v jiných se však ukazuje vztah mezi konkrétním významem a typem textu. V beletrii nacházíme téměř výhradně původní architektonický význam lemmatu *portál*, jehož se také týká zjištěný pokles, zatímco nárůst v odborné literatuře a publicistice je způsoben výrazně rostoucí frekvencí přeneseného významu označujícího portál internetový. Podobná situace nastává i u lemmatu *hejtman*, jehož původní historický význam typický hlavně pro beletrii je v odborné literatuře a publicistice zastiňován označením nově vzniklé funkce hlavního představitele kraje.

Soustředíme-li se nyní pouze na dvojice indikátorů v rámci jednoho hlavního typu textu, zjistíme, že v beletrii je nejčastějším průběhem právě dvojí nárůst, tj. nárůst mezi subkorpusy repre_bel_2000 a repre_bel_2005 i repre_bel_2005 a repre_bel_2010, který vykazuje celkem 42 výrazů. Podobná situace je také v publicistice se 45 případy dvojího nárůstu. V odborné literatuře naopak s celkem 53 výrazy ze 100 převažuje pokles mezi repre_odb_2000 a repre_odb_2005 následovaný nárůstem mezi repre_odb_2005 a repre_odb_2010. Tento zlom je nejčastěji doplněn dvojitým nárůstem v beletrii i publicistice, čímž se vracíme zpět k šestici ↑ ↑ ↓ ↑ ↑ ↑, která se ve výsledných tabulkách vyskytuje nejčastěji, a to v celkem 15 případech: *dětský hřiště*,

internet, *jestli*, *k vidění*, *m n*, *n m*, *opravdu*, *pár den*, *přehrávač*, *rok vězení*, *strašně moc*, *tenhle*, *určitě*, *vyrazit*, *za volant*.

Frekvenční průběh všech těchto výrazů v publicistických subkorpusech pub_RRRR a mf_RRRR (a tedy za použití 18 časových bodů) přitom lze charakterizovat jako nárůst, u řady z nich navíc výrazný a pravidelný. Lemmata *jestli* (obr. 6.5.16 na straně 161), *opravdu*, *tenhle*, *určitě*, *vyrazit* můžeme spolu s kombinacemi *k vidění*, *pár den*, *strašně moc* označit za výrazy typické pro neformální způsob vyjadřování. Ve výsledných tabulkách najdeme i další neformální výrazy s průběhem ↓ ↑ v odborné literatuře, avšak s jiným než ↑ ↑ v ostatních dvou hlavních typech textu, například *bavit*, *fotka*, *jenže*, *mobil*, *takhle*, *taky*, *teď*.

Tato zjištění naznačují, že se odborná literatura po roce 2000 – na rozdíl od publicistiky – nejprve formalizovala (jak zachycuje první pokles těchto neformálních výrazů) a vzápětí nastoupil opačný, ještě výraznější trend ke zvýšení neformálnosti vyjadřování; normalizovaná ARF v subkorpusu repre_odb_2010 totiž často výrazně převyšuje také normalizovanou ARF v subkorpusu repre_odb_2000. Překvapivé je přitom to, že právě subkorpus repre_odb_2005 se zdá být podle použitých jazykových prostředků formálnější než ostatní subkorpusy řady repre_odb_KKKK. Protože není příliš pravděpodobné, že by tato zjištění odpovídala jazykovému vývoji v odborné literatuře, budeme se složením těchto subkorpusů zabývat podrobněji.

Tabulka 6.2.6 ukazuje zastoupení vybraných žánrů ve výsledcích dotazu `[lemma="tenhle"]` na jednotlivé subkorpusy odborné literatury. Poznamenáváme, že u ostatních výše uvedených výrazů typických pro neformální vyjadřování byly výsledky podobné, lemma *tenhle* bylo vybráno jako jejich relativně frekventovaný a přitom žánrově nezávislý zástupce. Procenta v tabulce jsou vypočtena z uváděné frekvence, ne tedy normalizované ARF. Vybrány byly žánry s největším podílem výskytů v některém z uvedených subkorpusů, pro repre_odb_2000 je jím hudba (MUS), pro repre_odb_2005 společenský život (SCT) a pro repre_odb_2010 sport (SPO).

subkorpus	frekvence	MUS	SCT	SPO
repre_odb_2000	2 108	39 %	0 %	1 %
repre_odb_2005	1 970	4 %	40 %	0 %
repre_odb_2010	6 943	1 %	10 %	19 %

Tabulka 6.2.6: Zastoupení žánrů ve výsledcích dotazu `[lemma="tenhle"]`.

Je zřejmé, že takto radikální změny ve výskytu lemmatu *tenhle* v uvedených žánrech nemohou odpovídat skutečnosti, ale že jsou pravděpodobně způsobeny konkrétní skladbou textů. Zdůrazněme, že celkový počet rozlišovaných odborných žánrů se pohybuje okolo šedesáti, takže změny v řádech desítek procent jsou signifikantní, i

pokud vezmeme v úvahu nestejné zastoupení jednotlivých žánrů (viz tabulka 4.3.1 na straně 44, kde je také možné najít vysvětlení zkratek žánrů používaných v této podkapitole).

Podívejme se nyní blíže na složení uváděných žánrů ve vztahu k výsledkům v tabulce 6.2.6. V subkorpusu repre_odb_2000 je uvedený vysoký podíl hudby (MUS) v rozhodující míře způsoben zařazením časopisu Rock & Pop ročníku 1998, jehož celkový podíl na frekvenci lemmatu *tenhle* v subkorpusu repre_odb_2000 tvoří z celkových 2 108 výskytů plných 33 %; dalších 6 % doplňuje Folk & Country, ročník 1994. Podotýkáme přitom, že hudba je dostatečně zastoupena i v ostatních subkorpusech, avšak prostřednictvím formálnějších titulů, Rock & Pop v nich už nenajdeme. Podobná situace je i v ostatních subkorpusech, i když tady není monopolní postavení jednotlivých titulů tak patrné. Celkem 25 % výskytů lemmatu *tenhle* z celkových 1 970 v subkorpusu repre_odb_2005 pochází ze Story ročníků 2000 a 2001, které tak tvoří většinu kategorie společenský život (SCT); dalších 10 % přidává Xantypa ročníků 2003 a 2004. V subkorpusu repre_odb_2010 je pro výskyt lemmatu *tenhle* důležitý deník Sport ročníků 2007 až 2009, který tvoří 16 % z celkových 6 943 výskytů a dominuje tak kategorii sport (SPO).

Kromě zastoupení konkrétních titulů tady narážíme obecně na fakt, že proporce různých typů textu v rámci jednotlivých odborných žánrů nejsou nijak stanoveny. Je ovšem třeba podotknout, že kdyby tyto proporce stanoveny byly, nastal by v řadě oborů problém s jejich naplňováním. Skladba korpusu je tedy oportunistická v tom smyslu, že v rámci daných kategorií musí vycházet z textů, které jsou k dispozici v bance. Každou odbornou kategorii je přitom při výběru textů do reprezentativního korpusu potřeba naplnit především texty daného žánru, typ textu hraje při výběru textů spíše pomocnou roli (viz oddíl 4.2.6). Důsledkem však je složení odborných textů, které neumožňuje zjišťovat relevantní údaje o jazykových posunech nejenom uvnitř jednotlivých žánrů, ale jak se ukazuje, tak ani v odborné literatuře jako celku. Nepomůže přitom ani použití normalizované ARF, problémem je navíc měnící se zdroj neformálnosti a jeho rozsah, což jsou také hlavní důvody relativní formálnosti odborné literatury pozorované v subkorpusu repre_odb_2005.

Velká volnost je však možná i uvnitř kategorií jednoznačně daných dvojicí *txtype* a *genre*, příkladem mohou být „cestopisy, průvodce“ řazené do beletrie a určené dvojicí *txtype=FAC* a *genre=TRV*. Zatímco v SYN2000 je tato kategorie tvořena cestopisy, z nichž pouze jeden vyšel po roce 1989 a většina starších je dílem Karla Čapka, v SYN2005 převažují průvodce nakladatelství Olympia a v SYN2010 časopis Lidé a země. Typ textu se tedy mění de facto, ačkoli je stále označen jako FAC (literatura faktu). Problematické složení se tedy nevyhýbá ani beletrii a některé jeho důsledky jsou vidět i ve výsledných tabulkách: normalizovaná ARF zkratky *stol* v subkorpusu repre_bel_2005 je zhruba o řád vyšší než v ostatních beletristických subkorpusech

(viz tabulka 6.2.2 na straně 86, *cbf* rank 37), což je způsobeno především zmíněnými průvodci z nakladatelství Olympia, z nichž pochází 99 % výskytů této zkratky v celém subkorpusu repre_bel_2005.

Dalším problémem je celá kategorie 2. úrovně nazývaná „životní styl“, do níž jsou zařazeny mj. již zmíněné žánry SCT (společenský život) a SPO (sport); přesné složení této kategorie je vidět v tabulce 4.3.1 na straně 44. Jak již bylo uvedeno v oddílu 4.3.1, její zařazení do odborné literatury je přinejmenším sporné, protože obsahuje řadu populárních časopisů z oblasti domácích prací, zábavy, vztahů a událostí ze života celebrit, které nelze považovat ani za literaturu populárněnaučnou, a lze si jen obtížně představit jejich vědeckou obdobu. Praktické problémy při naplňování těchto kategorií a anotaci příslušných textů se potom projevují zejména tendencemi k interferenci s publicistikou; je například otázka, zda lze zmíněný deník Sport považovat za příklad populárněnaučného periodika se sportovním zaměřením.

Podobná situace nastává také u regionální publicistiky anotované jako *txtype=PUB* a *genre=REG*, tedy kombinací publicistického typu textu a odborného žánru z oblasti životního stylu (sic!). Publicistickým typem textu v kombinaci se žánrem z oblasti životního stylu byla původně anotována řada textů (zvláště staršího data z korpusu SYN2000), což ovšem ve spojení s nyní používaným principem kategorizace textů do hlavních typů textu (*txtype_group*) výhradně podle hodnoty *txtype* vede k tomu, že tyto texty jsou považovány za publicistické, a nemohou se tedy objevit v subkorpusech řady repre_odb_KKKK (k tomuto tématu viz „striktní *txtype*“ v oddílu 4.2.6).

Tento fakt vede spolu s občasnými opravami anotace starších textů (a s tím spojenými přesuny z odborné literatury do publicistiky a naopak) k tomu, že kategorie životní styl má v subkorpusech řady repre_odb_KKKK velice odlišné zastoupení. To je vidět v tabulce 6.2.7, která srovnává ideální zastoupení kategorie životní styl dané tabulkou 4.3.1 na straně 44 s reálným složením reprezentativních korpusů řady SYN. Zdůrazněme, že zvláště u korpusu SYN2000 jde o rozdíl způsobený změnou pohledu na zařazení kombinací publicistického typu textu s odborným žánrem; texty dané kategorie tedy v korpusu SYN2000 nechybějí, pouze se v korpusu SYN považují za publicistiku.

korpus	ideální	reálné
SYN2000	5,55 %	2,30 %
SYN2005	5,75 %	4,87 %
SYN2010	5,75 %	5,75 %

Tabulka 6.2.7: Zastoupení kategorie „životní styl“ v reprezentativních korpusech.

Tabulka 6.2.8 konkretizuje proměnlivé zastoupení jednotlivých typů textu v rámci kategorie životního stylu (vysvětlení zkratek je možné najít v tabulce 4.2.1 na straně 37). Je vidět jak různá celková velikost této kategorie v jednotlivých reprezen-

tativních subkorpusech, tak i její relativní formálnost v subkorpusu repre_odb_2005, která je dána vyšším podílem textů typu ENC (abecedně, systematicky a jinak uspořádaná díla). V subkorpusu repre_odb_2005 jsou jimi shodou okolností kuchařky naplňující žánr HOU („domácí práce, stravování, byt“), zatímco v ostatních subkorpusech je kuchařek relativně málo a výrazněji v nich převažují neformálnější časopisy.

txtype	repre_odb_2000	repre_odb_2005	repre_odb_2010
ADM	23 087	55 264	0
ENC	479 522	1 528 302	198 795
POP	1 795 926	3 149 554	5 413 417
SCI	0	83 198	123 896
TXB	0	54 822	13 696
celkem	2 298 535	4 871 140	5 749 804

Tabulka 6.2.8: Počty slov v jednotlivých typech textů kategorie „životní styl“.

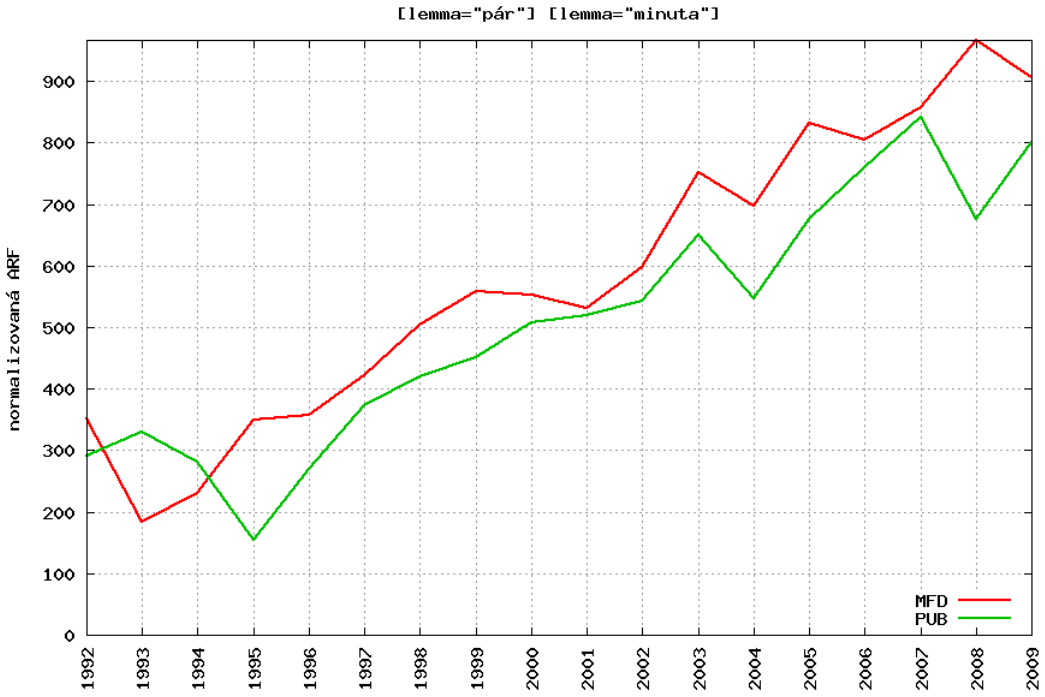
Závěrem shrnujeme problémy, na něž jsme v této podkapitole narazili. Jedním z nich je zařazení životního stylu do odborné literatury, což do ní vnáší řadu výrazů typu *celebrita*, *fotka*, *strašně moc* atd. Obecnějším problémem však je značná proměnlivost složení jednotlivých žánrů, v jejichž rámci je volba *txtype* dána především reálnými možnostmi banky, a je tedy do značné míry oportunistická. Výsledkem jsou příliš heterogenní data, což je pro diachronní srovnání blízkých stavů jazyka nevhodné. Ačkoli tím není nijak zpochybněna potřeba maximální různorodosti zdrojových textů (zvláště ze synchronního hlediska), nabízí se v této souvislosti otázka, jak odlišný odraz jazyka v synchronních reprezentativních korpusech bychom získali, pokud by jednotlivé kategorie byly naplněny podle stejných kritérií jako dosud, „pouze“ jinými texty s odlišným *txtype*.

I přes použití relativně velkého množství dat a jejich vyhodnocení pomocí ARF navíc nelze vyloučit vliv (ne)zařazení konkrétních textů, jak naznačuje příklad lemmatu *už* v tabulce 6.2.1 na straně 85: jeho dvojí pokles v beletrii (↓ ↓) je v rozporu s intuicí i zřejmým stabilním nárůstem v publicistice (obr. 6.2.2 na straně 96). Příčinu tohoto rozporu se zjistit nepodařilo, je však pravděpodobně důsledkem konkrétního složení beletristických subkorpusů. Vzhledem k těmto zjištěním je samozřejmě možná také opačná situace, kdy pozorovaný průběh sice odpovídá očekávanému, přesto však může být vzhledem ke složení dat a způsobu jejich vyhodnocení pouze výsledkem náhody, která skutečné posuny v jednotlivých hlavních typech textu jenom maskuje.

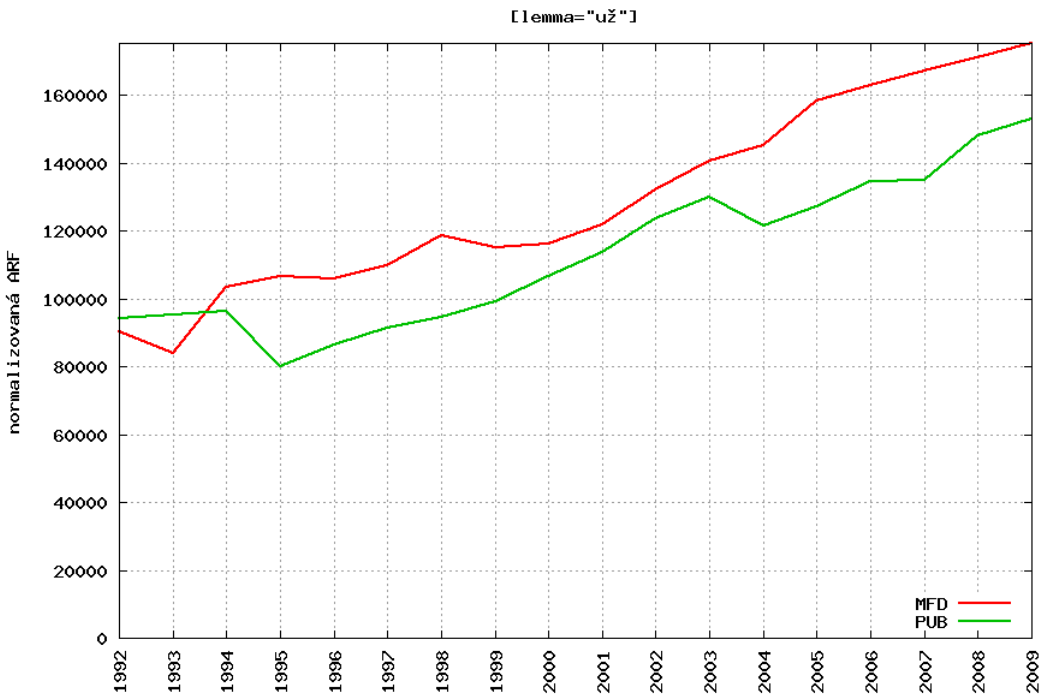
V neposlední řadě jde také o přílišnou hrubost numerické charakterizace frekvenčního průběhu jednotlivých výrazů, protože ze sledu pouhých tří hodnot (v rámci jednoho hlavního typu textu) nelze celkový trend spolehlivě určit. To je zásadní problém, pro-

tože není-li tento trend pravidelný (což většinou není), může ho malé množství časových bodů nejenom zastírat, ale dokonce s ním být v rozporu.

Řešením by tedy mohlo být nesnažit se zachytit posuny v celém psaném jazyce (byť vyhodnocované odděleně po jednotlivých hlavních typech textu), ale omezit se pouze na jeho menší a homogennější část, i když by toto omezení zmenšovalo rozsah dat a zvyšovalo tak vliv (ne)zařazení konkrétních textů. Bylo by ovšem potřeba zajistit více časových bodů, což naráží na nemožnost zjistit z dat rok vzniku textu, který v případě neperiodik často neodpovídá roku vydání. V jednotlivých reprezentativních subkorpusech se navíc roky vzniku textu beletrie a odborné literatury překrývají, což neumožňuje dostatečně spolehlivé vyhodnocení jazykového vývoje v nich. Jako metodologicky nejčistší se proto jeví sledování jazykového vývoje na publicistice, případně omezené na některý konkrétní titul, pro niž v zásadě platí, že rok vydání je zároveň také rokem vzniku textu. Publicistika tak poskytuje nejenom dostatečné množství časových bodů, ale zároveň by měla být z psaných typů textu nejvíce otevřená jazykových změnám.



Obrázek 6.2.1: Průběh normalizované ARF kombinace *pár minuta*.



Obrázek 6.2.2: Průběh normalizované ARF lemmatu *už*.

6.3 Iterativní *cbf*, *chi*, *ll* na publicistických subkorpusech

Tato podkapitola se zabývá rozborem výsledků iterativních metod *cbf*, *chi, ll* (viz oddíl 5.3.1) aplikovaných na publicistické subkorpusy řady pub_RRRR a mf_RRRR. Tyto subkorpusy jsou v záhlaví výsledných tabulek této a následujících podkapitol z prostorových důvodů uváděny pouze jako RRRR, například tedy 2006; konkrétní řada subkorpusů je vždy uvedena v popisku pod tabulkou. Jak již bylo zmíněno v oddílu 5.3.3, je pro iterativní *cbf*, *chi*, *ll* typické zvýrazňování velkých rozdílů ve sledovaných frekvencích. Protože jde při aplikaci na subkorpusy řady pub_RRRR a mf_RRRR o řadu 18 hodnot, znamená to především preferenci lemmat a kombinací s velkými výkyvy nebo oscilací.

Lexikální tabulky založené na pub_RRRR a mf_RRRR obsahují každá 50 lemmat, z nichž 20 najdeme v obou tabulkách (jde především o lemmata týkající se voleb). V tabulkách lexikálních kombinací je opakujících se položek 25, tj. celkem polovina; tematicky jsou velice podobné těm na lexikální úrovni. Frekvenční průběh všech výrazů můžeme rozdělit do tří hlavních skupin:

1. periodický průběh: jde o odraz pravidelně se opakujících událostí, jakými jsou např. volby nebo olympijské hry;

2. nepravidelný nárůst nebo pokles, případně spojený s oscilací: jde o odraz měnící se reality, ale také proměn publicistiky spočívajících především v pozvolném ředění původní politické orientace oddechovými tématy;

3. průběh s náhlými výkyvy (zpravidla směrem nahoru): jde o odraz nepravidelných, nečekaných událostí (významné zahraničně politické události a témata, povodně apod.), často však také o zvýraznění chyb v lemmatizaci nebo opomenutí při čištění korpusových textů.

Tato kategorizace samozřejmě není a nemůže být jednoznačná, často se objevují přechodové jevy spojené s interferencí několika příčin. U lemmatu $^{\mathrm{M}}$*lídr* (obr. 6.3.4) jde například o kombinaci periodicity a nárůstu, protože jde o neologismus spojený především s volbami. K interferencím dochází často také v případech, kdy má dané lemma více významů nebo různé kolokace s jiným frekvenčním průběhem, což může být patrné na úrovní lexikálních kombinací nebo při rozdělení celkové frekvence podle některé morfologické kategorie. Ilustrativním příkladem může být lemma $^{\mathrm{MP}}$*volba*, pod nějž jsou zahrnuty také *volby*, takže v plurálu zřejmou a výraznou periodicitu (obr. 6.3.1) v singuláru téměř nepozorujeme (obr. 6.3.2). Vzhledem k výrazné frekvenční převaze plurálu je frekvenční průběh celého lemmatu do značné míry dán průběhem

cbf	*chi*	*ll*	lemma	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	1	1	volba	20483	13725	17824	11946	23906	10676	20474	9821	11456	8543	17054	8572	8558	4834	10359	4433	12795	11423
2	4	3	volební	7453	4272	6011	3817	10265	3619	8740	3769	4930	2758	6462	2473	3197	1625	4235	1247	3553	3052
3	16	12	kandidátka	867	538	1459	573	1754	338	2451	794	1285	789	2534	571	1110	489	1762	373	1214	1570
4	18	17	předvolební	2591	1122	2181	1328	3901	939	3129	970	1193	934	2594	593	795	332	1254	264	1163	1342
5	5	5	foto	7442	12933	8234	3516	2079	1231	864	1404	410	501	1165	6242	5066	1014	767	3483	1631	1735
6	11	10	volič	5149	3717	4474	2820	6856	3088	5777	2678	3237	1962	4522	2024	2464	953	2407	678	2680	2457
7	7	6	povodeň	152	162	228	286	492	3346	3234	1492	1294	856	6440	4815	1904	1480	2239	1511	910	1469
8	117	108	povolební	607	98	387	138	715	173	949	199	176	165	568	59	86	33	505	84	354	139
9	84	81	práv	136	110	140	64	68	84	2081	2114	2283	1798	1682	1819	502	71	82	69	1163	1178
10	102	97	senátní	43	116	247	79	1728	512	1280	634	1172	497	1012	433	624	141	433	163	1214	393
11	37	33	komunální	938	1700	3490	1141	1301	1103	3496	1587	1375	1071	3019	1296	1466	1514	3420	1589	860	924
12	20	20	kandidát	6678	4434	5695	3847	8304	3957	6977	4427	6085	3463	6462	3687	3723	1942	3435	1931	4653	2875
13	73	69	kandidovat	1485	827	1692	840	2314	988	2321	1096	1405	921	2383	964	996	576	1290	388	1188	1317
14	39	38	irácký	1496	775	788	761	856	472	856	482	292	419	880	3283	1101	250	251	320	544	304
15	97	88	povodňový	38	46	42	64	162	1266	802	421	428	288	1966	1006	390	371	677	273	114	355
16	95	87	kosovský	141	139	42	20	34	65	954	2881	704	439	182	133	90	54	61	64	329	114
17	156	150	krimi	276	445	205	40	71	147	58	148	1157	1018	630	307	580	1177	1540	389	126	127
18	56	60	demokrat	3599	3318	4087	2682	5560	3366	5012	2896	2856	1965	3901	2469	2280	1322	2633	1028	3464	3065
19	53	53	koalice	5848	5625	5764	4691	6683	5048	4427	2607	2285	1649	4651	2826	2038	777	1553	946	2933	1735
20	26	23	přebor	206	301	191	64	503	596	1428	1402	1381	1217	1136	974	4415	8389	9269	9501	1492	1836
21	1064	1891	lisabonský	125	92	28	40	102	45	52	30	50	30	22	38	90	28	10	34	771	1266
22	45	39	sezona	27	179	1035	158	1712	3414	4542	5029	5290	5865	5925	5923	9980	15760	16673	16795	7953	8460
23	174	189	summit	1631	1393	1347	1546	1508	1756	863	886	1073	845	1913	876	292	109	81	131	834	1064
24	223	231	republikán	1594	1202	1626	1126	2576	1443	1933	665	713	352	545	273	333	60	100	78	620	418
25	96	96	záplava	770	711	629	835	1024	3310	2045	1100	1130	846	2680	1604	786	738	1066	664	733	849

Tabulka 6.3.1: Úroveň lexikální, *pub_RRRR*, *cbf chi ll*, 1. část.

cbf	*chi*	*ll*	lemma	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	151	167	prezidentský	3978	2781	2773	3185	3814	2092	2090	1668	1992	1166	1786	2169	1174	359	340	558	2769	1304
27	9	11	vláda	50300	40862	38161	40386	37218	35419	32300	26178	20613	18382	20207	17881	10800	5566	6135	5540	15994	16134
28	8	8	utkání	5979	6273	7605	3407	6453	5574	9414	9356	9566	9453	9080	8819	16919	26122	29088	28206	11025	11031
29	335	305	zaplavený	76	58	51	35	52	505	340	108	120	86	685	265	91	102	172	91	76	190
30	476	442	olympionik	504	121	196	69	346	137	312	114	375	153	214	110	385	207	293	137	468	101
31	6	7	zápas	9561	10262	11664	5427	10312	8672	13982	13490	14523	14633	14212	12962	22998	33138	36571	36005	17296	17008
32	222	242	měnový	2921	2099	1552	3674	2393	2735	2225	1381	2027	910	843	661	278	107	86	84	556	671
33	24	24	okresní	5512	5839	5573	3487	4272	4471	8354	9773	10697	10484	9327	4653	8200	12569	12069	11325	3755	3635
34	181	183	lidovec	233	445	1281	928	2741	2087	2736	1772	1501	1232	2342	1244	1224	723	1117	577	2074	1646
35	224	214	bombardování	412	526	499	484	298	147	218	1315	420	595	335	417	158	132	122	133	291	215
36	74	76	reforma	5995	5232	4352	4761	4212	4005	2831	2701	3033	2593	3219	4712	1892	812	757	1500	3173	1836
37	133	137	aliance	1588	2417	3509	3635	5089	4861	2461	3221	1391	1774	1728	889	421	151	185	227	708	595
38	15	16	hráč	7225	8811	11035	4637	9045	7752	12304	12404	12913	13130	12276	11439	18473	27549	28249	28074	13642	14184
39	27	30	politický	36852	26993	23952	20820	23799	17488	19375	14121	12044	11146	12444	9778	6403	3765	5211	3604	9761	10081
40	12	14	americký	22597	22905	25071	24164	25966	22226	19330	19701	19722	21050	18397	20003	11558	5826	5837	6931	16841	14475
41	32	35	miliarda	8244	9840	11529	23062	18521	15124	11513	10635	10640	10922	9958	10390	5664	3338	3689	4109	9217	11195
42	191	162	čs	14645	1757	704	311	262	230	369	288	248	184	260	222	196	136	122	108	266	190
43	161	157	federální	21166	6174	3132	3012	2704	1623	1271	1422	1414	1247	1027	831	467	197	161	154	607	823
44	22	22	branka	3610	3989	4795	2385	3992	3546	5994	5677	5623	5520	5508	4997	10673	17200	18461	17985	7839	8776
45	127	143	demokratický	12927	9724	9506	6331	8665	5824	5986	3784	3637	2779	3413	2380	1681	833	1163	680	2819	1988
46	144	168	akcie	3626	5932	10448	13457	10382	6968	5493	5489	4481	4004	3180	2472	1446	939	770	618	2819	1811
47	292	281	g	385	474	680	1738	359	659	436	330	377	330	215	271	376	455	434	1967	1365	975
48	155	149	teroristický	1165	1133	1202	919	1204	811	731	736	552	2410	2679	1802	1069	376	443	327	860	747
49	394	386	afghánský	607	405	200	306	306	212	237	206	155	880	812	355	173	59	65	142	670	405
50	172	172	koaliční	2081	3596	3751	2731	4584	3507	1745	1032	1022	697	1805	1649	1221	469	590	628	1580	811

Tabulka 6.3.2: Úroveň lexikální, *pub_RRRR*, *cbf chi ll*, 2. část.

cbf	*chi*	*ll*	lemma	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	1	1	foto	3040	1855	655	568	744	185	1557	3973	593	734	2238	23010	7205	1752	1318	1664	1548	1333
2	7	7	kandidátka	587	427	1428	786	1761	369	2998	713	1466	612	2948	425	1219	761	2510	589	1352	1496
3	3	3	volební	6270	3691	6346	3712	9871	3197	7779	2613	4039	1825	5609	1665	2886	2136	4666	1724	2931	2623
4	2	2	volba	18102	11908	18018	12446	21846	9094	16685	7021	9517	6040	13770	6041	8573	6767	13531	6535	10198	9371
5	15	14	předvolební	2038	742	2394	1223	4112	819	2615	646	931	527	2113	376	893	596	1840	553	1096	1160
6	9	9	povodeň	138	167	258	568	636	3831	4417	2026	1638	1064	6993	6111	2482	2217	3072	1935	1124	2090
7	12	12	volič	4128	3636	4832	2664	6903	2925	5105	1759	2457	1290	3588	1253	2392	1528	3116	1248	2322	2314
8	19	17	krátit	484	427	676	437	318	462	324	307	367	311	330	645	421	3751	501	368	359	573
9	29	24	kandidovat	1209	835	1632	1528	2650	952	2660	906	1349	736	2599	766	1106	854	1916	824	1272	1230
10	80	65	povolební	570	74	387	0	591	173	656	137	157	83	417	38	90	67	673	197	248	105
11	13	13	kandidát	6719	4749	5927	5546	8670	3693	6665	3318	5060	2557	5643	2618	3409	2263	3781	2341	3798	2314
12	20	19	komunální	1002	1873	3490	1310	1303	1391	4364	1920	1681	1274	3731	1510	1466	1229	3280	1367	1346	1371
13	30	30	koalice	6581	6418	6357	5371	5250	4391	2767	1465	1768	898	3213	1546	2027	1161	2400	1359	2430	1203
14	52	85	r	15857	16805	10738	10219	1081	889	659	998	1220	1122	2137	1108	958	721	650	701	787	591
15	169	182	čečenský	294	74	773	4367	1564	265	37	283	378	158	161	180	168	99	94	54	52	36
16	27	27	demokrat	2919	2801	3844	2664	5638	3399	4523	2014	2256	1354	2943	1503	2398	2098	4297	1972	3299	2540
17	11	10	krize	6495	5676	5154	7336	4640	4841	3795	3578	3070	3348	3243	2917	2113	2318	1997	2100	4283	15875
18	108	155	h	8343	10276	6797	6900	731	692	1417	1898	1226	1016	590	520	198	161	170	171	242	240
19	4	4	kraj	3144	3469	2330	1747	2314	2504	4614	5642	12269	24284	37577	42215	38468	48563	45785	42961	43310	46155
20	32	32	lídr	363	315	408	218	1754	1333	3646	1972	2826	2259	3872	1990	2641	2491	4839	2335	3660	3239
21	51	51	záplava	587	556	505	1659	1119	3202	1908	1009	922	767	2512	1765	861	1030	1435	906	549	858
22	35	34	tip	3265	4248	859	393	2352	2770	3792	4482	5279	5117	6398	6130	3042	7409	7732	6998	5541	5160
23	8	8	událost	8896	8495	5852	6288	4990	6324	6198	6853	7519	8100	7297	6578	7585	21363	28588	22951	20236	20559
24	22	22	galerie	674	4489	3071	2664	3839	6220	14516	17991	19627	21096	16117	13901	10712	5695	5736	6194	5769	5113
25	101	92	star	570	445	472	349	394	969	436	1642	779	305	896	293	324	359	334	305	327	309

Tabulka 6.3.3: Úroveň lexikální, *mf_RRRR*, *cbf chi ll*, 1. část.

cbf	*chi*	*ll*	lemma	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	734	1081	peruánský	432	111	107	1135	108	323	66	81	81	116	71	89	62	75	78	86	72	69
27	39	38	zajímavost	1900	1076	483	873	686	721	1082	1243	1448	1431	4450	3365	2070	4088	2256	1915	1865	1875
28	161	152	republikán	1676	1261	1911	2314	2447	1431	1451	397	418	195	421	110	314	143	227	151	347	170
29	72	67	lidovec	225	464	1342	1135	3146	2423	2232	1113	935	658	1606	732	1325	993	1909	1073	1617	1187
30	55	50	prsten	415	223	376	262	350	548	407	419	398	479	2683	2074	2007	583	405	368	310	327
31	47	45	grafika	1071	1558	494	262	1284	1991	5504	6465	6655	7388	5416	4618	3115	943	890	995	911	719
32	46	47	kampaň	3455	3023	4381	3494	6184	2706	3949	2055	2831	2031	3573	1985	2634	2291	3242	2301	3303	2948
33	61	55	rebel	207	352	397	393	572	490	497	626	597	2705	843	577	629	507	558	639	781	782
34	18	20	krátce	5959	4340	5809	6550	6134	6503	7699	7914	6357	4455	4604	6020	6898	14308	7929	6353	5248	5360
35	327	354	mm	864	612	1267	1616	197	190	183	709	168	108	188	168	183	174	166	193	146	150
36	79	76	off	605	1595	1868	2533	1786	1610	1932	2524	2887	3103	3228	919	206	178	126	153	190	184
37	21	21	okresní	8118	8811	7452	6812	5390	5666	12727	14144	16199	16542	14752	6270	4541	3852	3477	3350	3447	3216
38	16	15	výstava	5614	6455	4392	3450	5981	9884	24753	31155	32190	29885	23163	24014	18827	13244	12887	13543	12542	11714
39	236	255	škrt	259	223	172	480	470	1454	306	249	161	187	222	367	178	192	157	165	192	634
40	95	91	chřipka	276	371	558	786	763	1033	797	971	647	788	569	1026	687	1423	1528	941	651	2473
41	26	23	kino	2902	3524	2824	2314	2657	4755	11938	14330	14735	16367	15725	16708	14420	7149	6633	7351	5513	5523
42	357	345	olympionik	915	167	236	175	528	219	319	110	407	209	246	151	436	200	338	209	333	264
43	164	159	epizoda	1036	1076	462	568	350	381	375	1027	506	344	1008	325	285	523	353	384	435	309
44	157	175	e	6270	7920	6550	5939	1519	762	688	1191	1898	1130	436	238	193	151	130	133	110	161
45	97	98	str	4906	6826	5659	4367	3617	3145	1164	1005	521	844	2163	2087	1885	1225	1001	1437	2312	2585
46	98	135	únor	5407	6270	5369	18952	6045	5736	5536	5673	6013	5841	5476	5682	5885	6155	5974	6817	7307	7736
47	89	87	parlamentní	7428	6603	6464	6376	5129	2839	2559	1115	1029	1218	1955	796	810	991	1412	635	671	1080
48	90	83	federální	28536	7067	3071	3101	1920	1258	678	725	732	597	539	382	367	394	277	291	282	383
49	976	1179	bis	52	315	161	1266	267	231	205	211	132	228	163	38	71	69	59	80	96	56
50	134	136	koaliční	3040	4378	4231	3450	3973	3301	1151	621	705	429	1114	864	1155	739	942	796	1020	582

Tabulka 6.3.4: Úroveň lexikální, *mf_RRRR*, *cbf chi ll*, 2. část.

cbf	*chi*	*ll*	kombinace	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	1	1	komunální volba	428	983	2572	469	589	485	2081	754	486	257	1721	416	254	194	1718	342	228	241
2	3	3	volební kampaň	1165	532	909	494	1539	386	1310	524	639	386	1023	218	326	99	295	75	556	557
3	11	9	výsledek volba	911	428	783	296	1154	254	907	273	383	177	606	161	280	77	347	59	329	241
4	14	14	volební komise	455	185	704	316	1372	197	805	284	334	90	286	142	292	61	264	51	266	190
5	13	12	předvolební kampaň	802	399	829	528	1356	347	952	353	414	275	810	173	258	92	345	72	430	494
6	22	21	závislý kandidát	211	46	527	104	304	56	406	351	393	95	474	109	173	54	364	108	126	51
7	55	101	finanční krize	60	116	93	425	186	212	582	367	208	180	150	183	69	47	36	26	923	1874
8	16	18	prezidentský volba	1388	694	881	1422	1987	616	441	519	834	257	568	719	493	64	74	140	1315	405
9	34	38	volba prezident	976	243	84	69	147	332	391	155	152	118	476	523	121	38	54	67	695	190
10	70	73	prezidentský kandidát	401	156	186	207	361	80	89	120	247	35	171	207	121	11	7	52	620	63
11	7	8	reforma veřejný	0	104	168	114	84	102	80	247	435	512	619	1504	351	56	32	218	202	76
12	28	26	volební program	748	370	429	242	895	316	851	267	382	194	573	203	190	109	358	113	266	253
13	10	10	veřejný finance	27	46	37	104	105	152	183	140	155	356	368	1391	379	69	74	318	405	405
14	2	2	sociální demokrat	1285	1613	1999	1333	3780	2432	3948	2097	1808	1267	2517	1450	1238	867	1476	606	1934	1849
15	27	27	demokratický strana	2905	1873	2255	978	2327	1409	1727	914	885	503	855	424	407	151	393	136	695	405
16	35	32	občanský demokratický	1133	873	1519	879	1791	1038	1045	346	348	192	501	142	189	86	293	75	164	114
17	30	31	strana zelený	488	387	373	188	301	147	408	192	203	161	453	205	289	176	1151	571	1176	950
18	5	6	politický strana	4949	3792	3910	2968	4037	2206	4195	2449	2335	1918	2518	1293	1041	589	1135	438	1239	1900
19	19	19	parlamentní volba	2076	1896	2288	1956	2531	1612	1878	834	759	784	1465	479	308	224	440	140	531	747
20	17	15	teroristický útok	190	249	303	286	414	243	234	250	171	1515	1779	1098	665	250	290	179	417	380
21	58	59	koaliční strana	341	1029	1235	741	1586	1175	226	103	105	64	150	236	166	57	40	67	266	101
22	51	50	procento hlas	683	480	811	494	976	453	745	463	561	315	693	227	303	124	254	103	493	405
23	45	55	centrální banka	520	723	1021	2395	2529	2273	1872	1294	1322	936	872	813	307	119	98	93	797	861
24	15	17	sociální demokracie	1436	1359	1039	815	3110	2388	3386	1791	1477	1022	1740	1150	906	586	852	356	1264	1216
25	109	157	finanční trh	190	156	303	963	487	642	692	385	362	311	276	298	125	63	41	48	594	469

Tabulka 6.3.5: Úroveň lexikálních kombinací, *pub_RRRR*, *cbf chi ll*, 1. část.

cbf	*chi*	*ll*	kombinace	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	25	25	volební období	938	410	680	533	1225	583	1227	612	522	527	1210	608	602	585	1163	494	670	608
27	74	100	zahraniční politika	1626	1954	1333	998	1157	722	613	528	424	496	546	663	287	87	132	72	746	431
28	6	4	okresní úřad	2092	2434	2563	1743	2050	2053	3609	4132	4600	4257	3689	1083	571	460	357	241	152	63
29	32	42	miliarda dolar	1962	1914	2190	3116	2338	2029	1663	1675	1464	1508	1356	1573	697	280	214	249	1555	1507
30	63	61	předčasný volba	336	549	578	430	571	1010	1198	400	392	131	225	240	182	174	370	96	278	1076
31	145	167	pražský burza	43	306	1440	1032	1325	962	680	527	431	444	318	235	130	65	32	39	367	190
32	38	35	vládní koalice	2412	2145	2144	2030	2058	1628	858	553	466	285	744	1093	528	177	122	147	582	317
33	52	57	cenný papír	699	1653	3779	4682	3332	2740	1913	1485	1350	1251	990	748	378	195	101	116	683	431
34	99	90	komunální politika	33	92	210	49	92	58	407	159	181	125	344	139	178	178	329	121	76	63
35	138	133	hlasovací lístek	168	98	280	94	372	126	280	131	201	95	233	146	217	107	232	82	177	76
36	168	153	koaliční smlouva	38	81	112	49	175	72	173	101	118	42	243	68	85	33	103	64	190	101
37	209	264	hospodářský noviny	146	98	89	612	204	193	106	137	115	230	223	192	100	48	54	61	480	393
38	59	58	severoatlantický aliance	293	659	1160	1175	1469	1524	644	937	329	422	477	216	97	31	32	61	139	165
39	97	111	černý kronika	38	87	84	40	21	43	25	36	23	29	16	18	81	382	636	224	76	89
40	31	36	ministr zahraničí	4656	5018	4930	4114	3689	2857	2158	2123	1628	1803	1645	1674	780	278	289	302	1454	1266
41	37	34	olympijský hra	1984	1081	1076	598	1694	926	920	793	1435	868	910	772	1431	841	1091	813	1492	545
42	29	30	ministr obrana	2304	2850	3080	2469	2683	1988	1443	1312	913	1569	1470	1515	544	241	266	178	480	570
43	62	62	americký voják	298	543	536	227	432	186	140	207	230	321	470	927	396	113	87	152	354	266
44	86	80	kongresový centrum	43	35	42	25	466	889	501	235	624	345	485	352	230	128	206	219	455	329
45	77	86	národní banka	260	1775	1859	2375	2369	1805	1422	1087	1094	713	683	637	281	210	178	167	695	621
46	262	251	fotbalový mistrovství	76	58	196	35	251	74	165	74	99	63	113	52	106	83	192	69	164	63
47	169	222	francouzský prezident	434	318	359	548	539	472	223	168	229	181	225	306	114	39	34	72	531	557
48	175	175	ruský voják	271	410	401	652	516	100	61	298	306	126	93	108	53	58	41	61	190	76
49	18	16	evropský unie	764	688	5517	7966	6678	5212	5235	5703	6149	5739	6431	7753	7544	4966	4892	4095	5133	5661
50	8	11	spojený stát	4754	4417	5923	5803	6393	5022	3953	3827	3716	4909	4317	4534	2273	1204	1244	1328	3325	2989

Tabulka 6.3.6: Úroveň lexikálních kombinací, *pub_RRRR*, *cbf chi ll*, 2. část.

cbf	*chi*	*ll*	kombinace	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	1	1	komunální volba	432	1131	2588	830	731	652	2876	1030	741	344	2223	484	298	244	1922	541	341	412
2	9	9	výsledek volba	1002	297	848	393	1023	260	720	193	286	112	486	79	292	151	485	117	256	195
3	8	8	předvolební kampaň	708	185	891	655	1595	277	800	222	331	153	622	108	271	170	539	191	325	363
4	13	11	závislý kandidát	190	93	558	218	337	63	555	404	474	106	601	113	230	79	355	92	118	43
5	14	12	volební komise	397	223	999	393	1411	138	1018	258	307	54	225	72	172	71	159	54	116	76
6	7	7	volební kampaň	725	352	923	699	1754	467	1116	359	526	249	779	144	358	234	529	185	359	374
7	32	28	volební místnost	190	74	387	0	420	104	306	94	163	39	274	81	193	54	237	54	148	105
8	29	26	volební výsledek	415	241	365	131	636	110	441	126	161	56	257	42	135	115	313	72	168	130
9	47	66	bezpečnostní informační	363	1465	548	2795	375	231	101	87	54	93	38	49	47	30	27	62	42	52
10	23	23	volební program	622	204	365	262	947	387	821	200	333	147	498	138	167	168	413	163	246	217
11	3	3	sociální demokrat	1295	1744	2029	1485	4220	2747	3939	1682	1681	998	2077	1008	1380	1308	2377	1188	1849	1617
12	11	13	parlamentní volba	1935	2022	2459	1572	2199	1200	1257	428	363	415	1059	261	309	525	822	305	264	598
13	12	14	strana zelený	449	278	365	480	305	179	356	155	210	135	496	189	371	303	1611	918	1140	900
14	6	4	play off	501	1428	1804	2489	1659	1512	1860	2443	2831	3018	3042	800	124	63	67	78	110	112
15	19	20	občanský demokratický	1347	1465	2910	3232	2835	1569	1451	399	472	203	522	138	161	95	224	72	112	87
16	28	27	procento hlas	674	575	1020	437	1004	467	603	303	385	176	477	106	303	194	485	181	327	260
17	2	2	okresní úřad	3835	4062	3898	3494	2809	3001	5329	5794	6572	6341	5633	1569	623	361	287	255	196	155
18	31	29	finanční krize	69	204	118	655	248	254	367	258	154	141	135	198	80	63	44	54	621	1557
19	5	6	politický strana	4128	3506	3576	4192	3362	1812	3245	1561	1705	1274	1964	820	985	806	1511	697	1116	1299
20	33	32	předčasný volba	484	761	537	437	508	692	731	204	201	62	118	125	172	264	460	167	110	712
21	34	36	prezidentský volba	1641	612	848	1441	1729	519	207	213	398	112	280	395	457	198	212	374	617	296
22	18	19	demokratický strana	3006	1929	3178	2402	3025	1506	1669	671	772	338	745	263	329	246	401	159	290	318
23	65	57	volební zákon	311	315	762	218	680	115	332	231	376	178	177	32	86	40	132	26	36	60
24	49	43	objem obchod	35	798	2158	1878	1856	1067	303	408	47	33	21	15	9	26	15	14	18	25
25	10	10	výstavní síň	639	1410	472	306	394	825	2926	3760	4326	4273	2952	2474	1585	541	604	569	505	347

Tabulka 6.3.7: Úroveň lexikálních kombinací, *mf_RRRR*, *cbf chi ll*, 1. část.

cbf	*chi*	*ll*	kombinace	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	22	22	volební období	1002	297	655	611	1214	583	1249	550	439	465	1200	578	614	608	1047	611	701	614
27	81	129	konec leden	380	260	301	2533	464	433	340	444	445	421	426	429	449	505	451	535	499	506
28	56	60	koaliční strana	725	1428	1267	961	1773	1137	149	81	76	48	83	127	182	89	84	105	108	47
29	15	16	sociální demokracie	1278	1336	956	873	3451	2475	2971	1348	1237	728	1237	660	1028	1030	1452	728	1090	1039
30	41	40	měnový fond	898	556	354	1092	578	796	640	321	906	270	231	104	71	59	46	48	72	157
31	30	31	evropský parlament	1054	371	698	830	362	283	183	215	210	216	272	370	1162	511	365	269	306	822
32	64	62	pražský burza	35	649	2040	2664	1684	1448	532	363	128	135	81	43	86	87	107	115	196	99
33	43	41	mezinárodní měnový	898	594	354	1048	540	733	601	292	841	249	214	93	69	54	42	48	70	139
34	52	50	pravicový strana	622	334	365	262	375	231	840	256	177	116	282	100	79	77	180	68	94	152
35	67	77	informační služba	674	1818	698	3188	1163	825	462	368	282	201	169	170	180	170	237	422	395	414
36	21	21	olympijský hra	2539	1447	1203	1310	2275	1229	864	846	1497	857	909	832	1327	670	848	840	1368	571
37	108	102	koaliční dohoda	225	371	279	349	407	242	125	47	92	41	186	42	114	44	92	60	82	31
38	38	38	r o	3541	2819	1374	1223	311	323	258	516	412	394	1198	675	421	501	464	412	507	345
39	195	166	strana sociálně	0	352	365	306	572	167	159	38	87	31	49	6	21	10	21	18	18	13
40	39	34	teroristický útok	225	260	279	393	515	306	141	148	98	1027	1067	660	683	561	523	346	304	305
41	45	47	z přidaný	1036	1336	859	1790	534	739	391	242	235	203	205	350	910	365	353	324	280	354
42	16	15	krajský úřad	155	148	290	306	261	179	319	366	855	3101	3311	4425	4320	4112	4091	3629	3363	3575
43	4	5	volný čas	501	594	580	437	756	894	1175	1398	1553	1695	3427	5068	2422	1647	2218	2205	1939	2047
44	57	53	školský úřad	812	1076	526	175	566	779	896	983	1043	197	94	40	22	16	13	20	20	4
45	46	48	ministr zdravotnictví	829	1187	1353	1485	1475	1229	465	702	443	328	205	136	253	281	707	539	627	235
46	54	45	přednosta okresní	656	761	698	524	483	467	755	850	996	566	674	68	36	12	19	20	10	13
47	191	183	izraelský voják	173	519	612	655	280	202	50	29	119	122	131	53	28	28	90	14	22	40
48	100	90	americký seriál	535	334	698	437	1360	242	74	101	58	33	36	28	45	24	31	84	76	29
49	62	64	severoatlantický aliance	397	909	1546	1135	1551	1552	436	520	188	232	252	104	101	61	84	109	98	72
50	20	18	městský divadlo	311	835	419	393	331	606	2581	3412	3547	3837	3378	3586	3113	1542	1653	1660	1460	1109

Tabulka 6.3.8: Úroveň lexikálních kombinací, *mf_RRRR*, *cbf chi ll*, 2. část.

plurálu, mírná oscilace v singuláru je způsobena někdy nesprávným určením pádu a čísla u homonymního tvaru *volby*.

Mezi lemmaty a kombinacemi 1. skupiny s periodickým průběhem je na první pohled výrazný podíl výrazů týkajících se voleb. Při podrobnějším zkoumání zjistíme, že jde až na výjimky o volby v ČR a že jsou tyto výrazy v tabulkách většinou na předních místech. Velice podobný je také jejich frekvenční průběh, který se vyznačuje velkou, pravidelnou oscilací odrážející závislost jejich frekvence na čase. Lemmaty a kombinacemi tohoto typu s velmi výraznou volební periodicitou jsou $^{\mathrm{P}}$*hlasovací lístek*, $^{\mathrm{M}}$*kampaň*, $^{\mathrm{MP}}$*kandidát*, $^{\mathrm{MP}}$*kandidátka*, $^{\mathrm{MP}}$*kandidovat*, $^{\mathrm{P}}$*koaliční smlouva*, $^{\mathrm{MP}}$*komunální*, $^{\mathrm{P}}$*komunální politika*, $^{\mathrm{MP}}$*komunální volba*, $^{\mathrm{MP}}$*povolební*, $^{\mathrm{MP}}$*předvolební*, $^{\mathrm{MP}}$*předvolební kampaň*, $^{\mathrm{P}}$*senátní*, $^{\mathrm{MP}}$*volba*, $^{\mathrm{MP}}$*volební*, $^{\mathrm{MP}}$*volební kampaň*, $^{\mathrm{MP}}$*volební komise*, $^{\mathrm{M}}$*volební místnost*, $^{\mathrm{MP}}$*volební období*, $^{\mathrm{MP}}$*volební program*, $^{\mathrm{M}}$*volební výsledek*, $^{\mathrm{MP}}$*volič*, $^{\mathrm{MP}}$*výsledek volba*, $^{\mathrm{MP}}$*závislý kandidát*. Tato jejich periodicita se jasně projeví, vyneseme-li čísla z tabulek do grafu. Na základě frekvenčního průběhu adjektivního lemmatu $^{\mathrm{MP}}$*komunální* na obr. 6.3.3 tak například můžeme usoudit, že komunální volby se konaly od roku 1994 každé 4 roky, což se také shoduje se skutečností.

Ne všechna lemmata patřící do 1. skupiny však mají takto zřetelný průběh, což je dáno větší růzností událostí nebo osob, k nimž se dané lemma vztahuje, stejně jako větší růzností kolokátů. Například (občanský nebo sociální) $^{\mathrm{MP}}$*demokrat* nebo $^{\mathrm{MP}}$*lidovec* sice většinou označují členy českých politických stran, ale nejenom je; podobně se i adjektivum $^{\mathrm{P}}$*prezidentský* týká také prezidentské kanceláře či úřadu, nejenom voleb ve Spojených státech nebo Rusku. Těmto zahraničním událostem odpovídají kombinace $^{\mathrm{P}}$*prezidentský kandidát* a $^{\mathrm{MP}}$*prezidentský volba*, zatímco $^{\mathrm{P}}$*volba prezident* má zcela jiný průběh a týká se (nepřímé) volby prezidenta ČR, případně ČSFR. Také frekvenční špičky kombinace $^{\mathrm{M}}$*evropský parlament* v zásadě odpovídají volbám do Evropského parlamentu, v roce 1999 však výraznější špička chybí a celý frekvenční průběh tak připomíná spíše oscilaci.

Časté je také spojení periodicity s celkovým nárůstem nebo poklesem. Příkladem periodického nárůstu je již zmiňované lemma $^{\mathrm{M}}$*lídr* (obr. 6.3.4) a kombinace $^{\mathrm{MP}}$*strana zelený*, u níž je patrné výrazné zvýšení periody v roce 2006, kdy se Strana zelených dostala do parlamentu. Periodickým poklesem se naopak vyznačují lemmata a kombinace $^{\mathrm{MP}}$*demokratický strana* (ODS), $^{\mathrm{MP}}$*koalice*, $^{\mathrm{M}}$*koaliční dohoda*, $^{\mathrm{MP}}$*občanský demokratický* (ODS, dříve i ODA), $^{\mathrm{M}}$*parlamentní*, $^{\mathrm{MP}}$*parlamentní volba*, $^{\mathrm{M}}$*pravicový strana*, $^{\mathrm{MP}}$*republikán*, $^{\mathrm{MP}}$*sociální demokracie*, $^{\mathrm{MP}}$*sociální demokrat*, $^{\mathrm{M}}$*strana sociálně* (ČSSD), $^{\mathrm{M}}$*volební zákon*. Některé z uvedených kombinací jsou částmi plných názvů českých politických stran, jednou z příčin poklesu je v jejich případě stále více převažující způsob vyjadřování pomocí zkratek.

Do 1. skupiny patří také lemma $^{\mathrm{MP}}$*olympionik* a kombinace $^{\mathrm{MP}}$*olympijský hra* s pravidelně se střídajícími nižšími a vyššími maximy odpovídajícími rokům konání

zimních a letních olympijských her. Podobného periodického typu je také kombinace $^{\mathrm{P}}$*fotbalový mistrovství* odkazující na mistrovství světa nebo Evropy ve fotbale.

Do 2. skupiny jsme zařadili výrazy, které vykazují (většinou nepravidelný) pokles; o nárůst jde spíše výjimečně. Tato skupina je menší a různorodější než první, patří do ní především lemmata a kombinace spojené s (často zahraniční) politikou a ekonomikou $^{\mathrm{P}}$*americký*, $^{\mathrm{P}}$*demokratický*, $^{\mathrm{MP}}$*koaliční*,[3] $^{\mathrm{MP}}$*koaliční strana*, $^{\mathrm{P}}$*měnový*,[4] $^{\mathrm{P}}$*miliarda*, $^{\mathrm{P}}$*miliarda dolar*, $^{\mathrm{P}}$*ministr obrana*, $^{\mathrm{P}}$*ministr zahraničí*, $^{\mathrm{M}}$*ministr zdravotnictví*, $^{\mathrm{P}}$*politický* (obr. 6.3.5), $^{\mathrm{P}}$*reforma*, $^{\mathrm{MP}}$*severoatlantický aliance*, $^{\mathrm{P}}$*spojený stát* (Spojené státy), $^{\mathrm{P}}$*summit*, $^{\mathrm{P}}$*vláda*, $^{\mathrm{P}}$*vládní koalice*, $^{\mathrm{P}}$*zahraniční politika* (obr. 6.3.6), jejichž více či méně hladký frekvenční pokles je důsledkem klesající politické a ekonomické orientace publicistiky, do níž tak stále více pronikají nepolitická, oddechová témata.

Další společný rys mají frekvenční průběhy lemmat a kombinací $^{\mathrm{P}}$*akcie*, $^{\mathrm{P}}$*cenný papír*, $^{\mathrm{P}}$*centrální banka*, $^{\mathrm{P}}$*finanční trh*, $^{\mathrm{P}}$*národní banka* (většinou jde o Českou národní banku), $^{\mathrm{M}}$*objem obchod*, $^{\mathrm{MP}}$*pražský burza*, a to výrazné maximum v letech 1993–1997. To odpovídá obnovení Burzy cenných papírů a zavedení RM-systému v roce 1993, a hlavně výraznému nárůstu rozsahu jednoho čísla MFD v bance právě v letech 1997 a 1998 zmíněnému už v podkapitole 4.5 (viz také obr. 4.5.5 na straně 59). Jeho příčinou je řada nových zájmových příloh, která způsobila skokové „naředěníˇ původní politické a ekonomické orientace a tím i strmější frekvenční propad řady výrazů spojených s politikou a ekonomikou právě v tomto období, což se týká zvláště lemmat a kombinací $^{\mathrm{P}}$*akcie*, $^{\mathrm{MP}}$*koaliční*, $^{\mathrm{P}}$*národní banka*, $^{\mathrm{M}}$*parlamentní*, $^{\mathrm{MP}}$*republikán*, $^{\mathrm{MP}}$*severoatlantický aliance*, $^{\mathrm{P}}$*vládní koalice*.

Poznamenejme přitom, že synonyma jako $^{\mathrm{P}}$*akcie* a $^{\mathrm{P}}$*cenný papír* mají velice podobný průběh, takže nejde o pouhou záměnu jednoho pojmu jiným, ale o skutečný pokles zájmu o toto téma, nebo spíše o zvýšený zájem tisku o dění na burze v polovině 90. let jako důsledek kupónové privatizace. Podobné případy, kde stejné tendence projevuje používání plného názvu a jeho zkratky, jako například *Bezpečnostní informační služba* a *BIS* nebo *Severoatlantická aliance* a *NATO*, jsou poměrně časté. Někdy však lze vysledovat mírně odlišný vývoj, například *Mezinárodní měnový fond* vs. *MMF*, nebo přímo sílící tendenci k užívání jednoho z prostředků, konkrétně *DPH* namísto *daň z přidané hodnoty*.

Velice zřetelný pokles zaznamenala hned po roce 1992 lemmata $^{\mathrm{P}}$*čs* a $^{\mathrm{MP}}$*federální*, což je samozřejmě dáno rozpadem ČSFR. Změna politické reality je také rozhodující příčinou frekvenčního nárůstu lemmatu $^{\mathrm{M}}$*kraj* a kombinace $^{\mathrm{M}}$*krajský úřad*. Průběh frekvence lemmatu $^{\mathrm{MP}}$*okresní*, který vidíme na obr. 6.3.7, je ovlivněn několika různými

[3] Pro frekvenční průběh lemmatu $^{\mathrm{MP}}$*koaliční* je rozhodující frekventovaná kombinace $^{\mathrm{MP}}$*koaliční strana*, která vykazuje v letech 1997 a 1998 strmý pokles a u které není patrná žádná periodicita. Vliv této kombinace tak převáží vliv ostatních, například silně periodické, ale málo frekventované kombinace $^{\mathrm{P}}$*koaliční smlouva*.

[4] Často jde o součást spojení *Mezinárodní měnový fond*, které svým průběhem patří do 3. skupiny.

faktory. Jeho nárůst po roce 1997 je způsoben již zmíněným rostoucím podílem regionálních mutací MFD, hlavním důvodem poklesu po roce 2002 je však historický vývoj, konkrétně zánik okresních úřadů ke konci roku 2002, což lze dokumentovat nápadnou shodou grafu lemmatu $^{\mathrm{MP}}$*okresní* s grafem frekventované kombinace $^{\mathrm{MP}}$*okresní úřad* na obr. 6.3.8. Výjimkou z této shody je publicistika z let 2004–2007, kde dochází k opětovnému strmému nárůstu frekvence lemmatu $^{\mathrm{MP}}$*okresní*, který neodpovídá ani frekvenci kombinace $^{\mathrm{MP}}$*okresní úřad*, ani průběhu v MFD. Jeho příčinou je výrazná převaha regionálních titulů VLP v subkorpusech pub_2005, pub_2006, pub_2007 (a částečně také pub_2004) popsaná v podkapitole 4.5 (viz také obr. 4.5.4 na straně 58) a velký podíl sportovního zpravodajství v nich. To dokazují typické kolokace v publicistice z těchto let, kterými jsou *přebor*, *soutěž*, *kolo*, zatímco v předchozích letech šlo především o *úřad*, *soud*, *město*. Zánik okresních úřadů je také hlavním důvodem výrazného snížení frekvence kombinace $^{\mathrm{M}}$*přednosta okresní*, která bývá součástí širšího spojení *přednosta okresního úřadu*, stejně jako je zánik školských úřadů příčinou obdobného poklesu u kombinace $^{\mathrm{M}}$*školský úřad*.

Pomyslný přechod mezi 2. a 3. skupinou tvoří lemmata a kombinace, jejichž frekvence sice klesá, tento pokles je ale nepravidelný a narušovaný řadou více či méně výrazných výkyvů. Sem lze zařadit kombinaci $^{\mathrm{P}}$*r o*, která je součástí širší *spol. s r. o.* nebo *s. r. o.* Také ona vykazuje výrazný pokles, jeho nerovnoměrnost je přitom způsobena řadou faktorů, od kupónové privatizace přes proložené písmo v textech až po uvedení filmu Příšerky s. r. o. v roce 2002. Podobným způsobem klesá i kombinace $^{\mathrm{M}}$*z přidaný* (součást spojení *daň z přidané hodnoty*), v tomto případě však jde o projev tendence k nahrazení celého spojení zkratkou DPH, která sice vykazuje podobné výkyvy, ale spojené naopak s celkovým růstem.

Také kombinace $^{\mathrm{M}}$*mezinárodní měnový* a $^{\mathrm{M}}$*měnový fond* (součásti spojení *Mezinárodní měnový fond*) vykazují pokles spojený s výkyvy způsobenými řadou vlivů, například zasedáním MMF a Světové banky v Praze v roce 2000. Kombinace $^{\mathrm{P}}$*kongresový centrum* naopak spíše roste, na výkyvech a celkovém nárůstu se podepsalo především přejmenování Paláce kultury na Kongresové centrum Praha, dále významné události v něm konané (zasedání MMF a Světové banky v roce 2000, zasedání NATO v roce 2002), diskuse o jeho rekonstrukci a kulturní program v MFD zejména v roce 1997.

Frekvenční průběh kombinací $^{\mathrm{P}}$*americký voják*, $^{\mathrm{M}}$*izraelský voják*, $^{\mathrm{P}}$*ruský voják* osciluje podle aktuálních zahraničněpolitických událostí, podobně jako v případě kombinace $^{\mathrm{P}}$*francouzský prezident*; patrná je však opět tendence k celkovému poklesu daná naředěním publicistiky a odklonem od zahraničních témat k domácím. Jednou z mála rostoucích kombinací je $^{\mathrm{P}}$*evropský unie*, jejíž frekvenční špička v letech 2003–2004 odráží vrcholící diskuse o vstupu ČR do EU v době podepsání Smlouvy o přistoupení a vlastního vstupu do EU, naopak velice nízká frekvence této kombinace v letech 1992–1993 je dána tím, že Evropská unie nahradila Evropské společenství až v těchto letech.

Dostáváme se tak ke 3. skupině výrazů vyznačujících se průběhem s výraznými výkyvy. Její jasně vymezenou částí je čtveřice lemmat [MP]*povodeň*, [P]*povodňový*, [MP]*záplava*, [P]*zaplavený*, jejichž dva výrazné výkyvy v letech 1997 a 2002 zřetelně ukazují na dvě největší povodně ve sledovaném období u nás. Podobného charakteru je také sezónní výskyt lemmatu [M]*chřipka* s maximy v letech 2006 (*ptačí chřipka*) a 2009 (*prasečí chřipka*), kombinace [MP]*předčasný volba*, jejíž frekvence závisí především na tom, jsou-li předčasné volby v dané době politickým tématem, a lemmatu [M]*škrt* s výrazným maximem v roce 1997 (*rozpočtové škrty*). Frekvence kombinací [P]*reforma veřejný* a [P]*veřejný finance* (součásti spojení *reforma veřejných financí*) kulminuje ve sledovaných subkorpusech v roce 2003, zatímco lemma [M]*krize* a kombinace [MP]*finanční krize* v roce 2009. Výraznou sezónností se vyznačují také frekvenční průběhy některých lemmat, které jsou součástí názvu nebo podtitulu filmu: [M]*epizoda* (Star Wars), [M]*prsten* (Pán prstenů), [M]*rebel* (muzikál Rebelové), [M]*star* (Star Wars a Star Trek).

Další část 3. skupiny tvoří výrazy [P]*afghánský*, [P]*bombardování*, [M]*čečenský*, [P]*irácký*, [P]*kosovský*, [P]*lisabonský*, [M]*peruánský*, [P]*teroristický*, [MP]*teroristický útok* odkazující na významné zahraničněpolitické události; z čísel ve výsledných tabulkách lze opět snadno vysledovat roky, v nichž se tyto události odehrály. Přítomnost mnoha depropriálních adjektiv tady ukazuje na nezbytnost vyřazení proprií v průběhu zpracování, v opačném případě bychom nacházeli především je. K těmto výrazům lze zařadit také lemma [P]*aliance*, jehož frekvenční oscilace je dána převážně intenzitou diskusí o Severoatlantické alianci.

Specifickou část 3. skupiny tvoří lemmata spojená se sportem, která se vyskytují pouze v tabulce založené na subkorpusech pub_RRRR: [P]*branka*, [P]*hráč*, [P]*přebor*, [P]*sezona*,[5] [P]*utkání*, [P]*zápas*. Její specifičnost spočívá v tom, že frekvenční průběh všech těchto lemmat je velice podobný lemmatu [P]*utkání* s výrazným vrcholem v letech 2005–2007 (obr. 6.3.9). Tento vrchol se však týká jenom publicistiky, nikoli MFD, a je opět způsoben již zmíněnou výraznou převahou regionálních titulů VLP v publicistice těchto let. Podíl sportu v titulech VLP je zjevně nadprůměrný, jejich regionálnost potvrzuje frekvenční průběh lemmatu [P]*přebor*, jehož nárůst je ze všech lemmat této skupiny největší. Nejde však jen o sport, vliv VLP na výsledky za toto období a jeho orientace na domácí, nepříliš politická nebo ekonomická témata je vidět na propadu řady lemmat a kombinací v témže období: [P]*americký*, [P]*centrální banka*, [P]*francouzský prezident*, [P]*miliarda*, [P]*vláda*, [P]*zahraniční politika* (obr. 6.3.6) a řady dalších. Tyto příklady prakticky dokazují, že složení subkorpusů řady pub_RRRR vybraných jako publicistické subkorpusy korpusu SYN je zejména v letech 2005–2007 nevhodné, a nelze je proto považovat za reprezentativní zástupce publicistiky.

[5] Kvůli nevhodné lemmatizaci interferuje s paralelně existujícím lemmatem *sezóna*, pod nějž jsou zahrnuty i některé tvary s krátkým *o*.

Naopak pouze v tabulce založené na subkorpusech mf_RRRR se vyskytují lemmata a kombinace spojené s uměním: [M]*galerie*, [M]*grafika*, [M]*kino*, [M]*městský divadlo*, [M]*výstava*, [M]*výstavní síň*. Kulminace jejich výskytu v letech 1998–2004 pochází z kulturních programů MFD, a je tedy způsobená složením dat, konkrétně přítomností dané rubriky v nich. Stejnou příčinu, tedy složení dat, mají také výkyvy kombinace [M]*americký seriál*, lemmatu [M]*off* a kombinace [M]*play off*.

Poslední, ale významnou část 3. skupiny lemmat s výraznými výkyvy tvoří chyby způsobené nedostatečným čištěním nebo špatnou lemmatizací. Většina z nich se vyskytuje pouze v tabulkách založených na subkorpusech řady mf_RRRR, což je dáno tím, že v publicistice jako celku se chyby vztahující se k jedinému periodiku výrazněji neprojevují. Důsledkem nedostatečného čištění názvů rubrik jsou výkyvy lemmat [M]*krátce*, [P]*krimi*, [M]*událost* (Události), [M]*tip* (Kulturní tipy, Tipy na víkend, Tipy pro volný čas apod.), [M]*zajímavost* (Zajímavost) a kombinace [P]*černý kronika*,[6] popisky pod obrázky jsou zdrojem výkyvů u lemmatu [MP]*foto*, odkazy na další strany u lemmatu [M]*str*. Podobně lemma [M]*krátit* se objevuje téměř výhradně jako část dovětku „Dopisy jsou redakčně kráceny“, který nedopatřením nebyl smazán z MFD ročníku 2005. Poněkud odlišný je případ kombinace [M]*volný čas* (obr. 6.3.10): ačkoli je výkyv v MFD ročníku 2003 způsobený nedostatečným čištěním názvu rubriky „Tipy pro volný čas“, je zřejmý pozvolný celkový frekvenční nárůst této kombinace v MFD i publicistice, který v letech 1992 a 2009 dosáhl téměř čtyřnásobku a který tak demonstruje vzrůstající oblibu oddechových témat v publicistice.

Jako [P]*práv* jsou lemmatizovány převážně tvary *práva* a *právu* omylem považované za jmenný tvar adjektiva; ve většině případů jde navíc o výskyty z deníku Právo, typicky v obměnách spojení „řekl(a) Právu“. Vzhledem k tomu, že skutečný výskyt lemmatu *práv* v datech je velice vzácný, jeho frekvenční průběh ukazuje především zastoupení tohoto deníku v subkorpusech řady pub_RRRR (a tedy také v korpusu SYN), které začíná v roce 1998 a které je v letech 2005–2007 velice omezené, což je opět dáno tím, že subkorpusy pub_2005, pub_2006 a pub_2007 obsahují minimum textů mimo produkci VLP. Podobné povahy je také kombinace [P]*hospodářský noviny*, která se ovšem kromě Hospodářských novin vyskytuje poměrně často i v jiných periodikách a jejíž výkyvy nejsou způsobeny chybnou lemmatizací. Za nedostatek použité lemmatizace lze považovat také nerozpoznání zkratky BIS, která byla lemmatizována jako [M]*bis* (s nerozpoznaným slovním druhem) a dostala se tak nedopatřením k apelativům ve výsledných tabulkách.

Jiný typ chyby se vyskytl u lemmat [M]*e*, [M]*r*, [M]*h*; je jí nerozpoznání velkého množství nadpisů psaných proloženým písmem a týká se pouze starších textů do roku 1995. Toto vysvětlení platí u [P]*g* pouze v omezené míře, protože jde převážně o označení

[6] Tady jde o nedůsledné čištění VLP, projevuje se však na celku publicistiky, protože VLP tvoří v letech 2005–2007 její výraznou část.

gramu v kuchařských receptech. Jeho výkyv v roce 1995 je dán tím, že celému tomuto ročníku týdeníku Vlasta (obsahujícího množství receptů) byl přiřazen *txtype=PUB*, takže Vlasta ročníku 1995 byla zařazena do publicistiky.

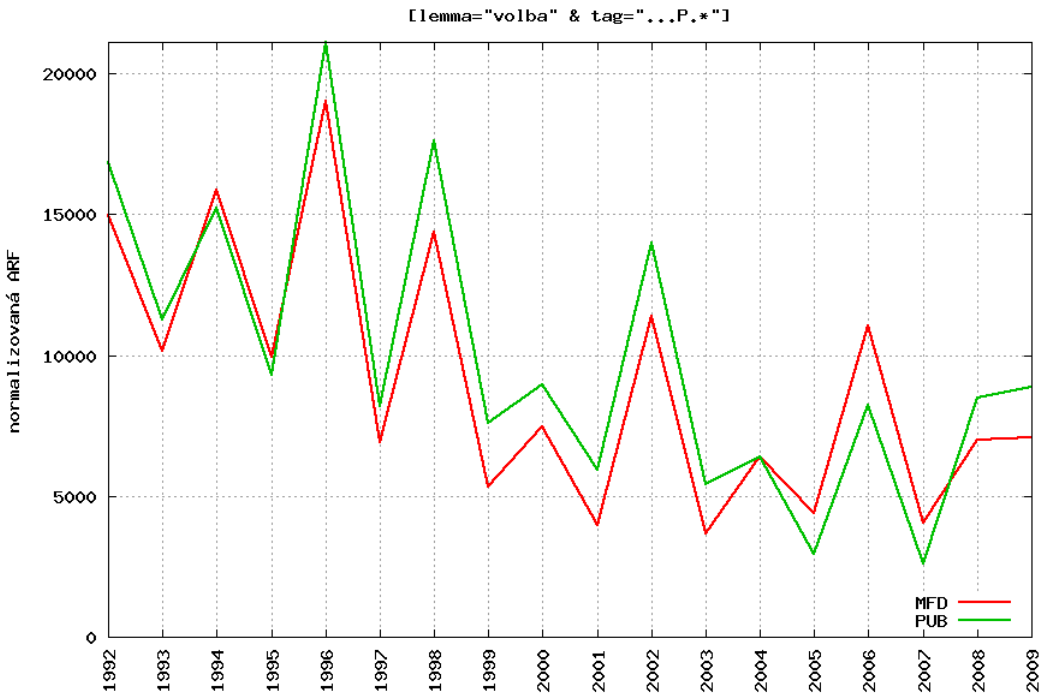
Několika různými faktory jsou způsobeny výkyvy lemmatu $^{\mathrm{M}}$*mm*; jde jednak o novinářskou zkratku nevyčištěnou z MFD let 1994 a 1995, ale také o označení milimetru v seriálu o fotoaparátech a objektivech v MFD ze začátku února 1995. Tato koncentrace označení $^{\mathrm{M}}$*mm* pouze v několika málo číslech by byla za normálních okolností odfiltrována pomocí ARF, protože se však MFD z tohoto roku skládá pouze z lednových a únorových čísel (jde tedy o pouhý fragment celého ročníku, viz obr. 4.5.5 na straně 59), nebyla frekvence $^{\mathrm{M}}$*mm* redukována dostatečně, a dostala se proto i do výsledných tabulek. Tutéž příčinu, tedy koncentraci určitého tématu do MFD z ledna a února 1995, mají také výrazné výkyvy lemmat $^{\mathrm{M}}$*bis*, $^{\mathrm{M}}$*únor* a kombinací $^{\mathrm{M}}$*bezpečnostní informační*, $^{\mathrm{M}}$*informační služba*, $^{\mathrm{M}}$*konec leden*.

Závěrem shrňme, že ačkoli je z úvodních tabulek a následné diskuse vidět, o čem se v publicistice psalo a o čem nikoli, a tím také zprostředkovaně o některých historických událostech, nedozvěděli jsme se téměř nic o jazyce a jazykovém vývoji. To je dáno do značné míry použitou metodou, přesto však výsledky nepovažujeme za nepodstatné nebo banální; jsou totiž cennou informací o povaze dat a o vlivu, jaký mohou mít rozdíly ve složení použitých subkorpusů spolu s chybami při jejich zpracování na celkové výsledky.

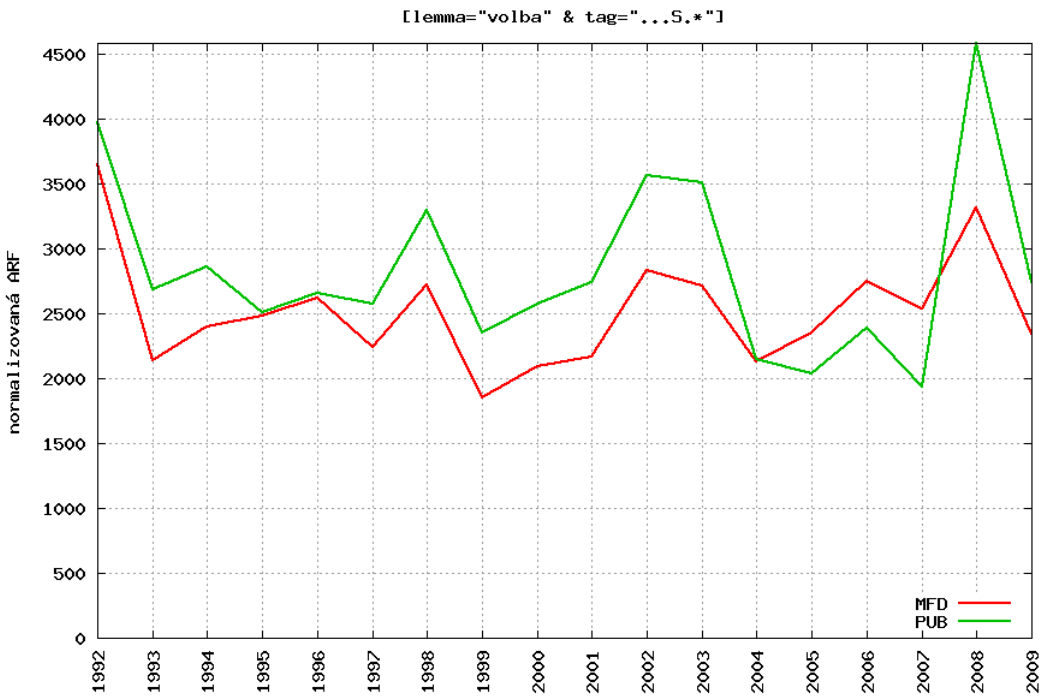
Z řady průběhových grafů je v letech 1992–1997 zřejmá – kromě důsledků neúplnosti ročníku 1995 – také celková rozkolísanost frekvenčního průběhu v tomto období. Ta je způsobena neúplností starších dat spolu s jejich menším množstvím a neustáleným způsobem čištění v prvních fázích budování korpusu v 90. letech (viz oddíl 4.2.5), a týká se tedy jak subkorpusů řady mf_RRRR, tak i pub_RRRR. V letech 1998–2003 je již složení MFD a publicistiky v zásadě ustálené a s kompletními ročníky (v této době také přibyly regionální přílohy a magazíny), takže i frekvenční průběhy jsou víceméně stálé a vyznačují se také poměrně stabilním rozdílem mezi MFD a celkem publicistiky. Od roku 2004, zvláště v letech 2005–2007, se však v publicistických subkorpusech projevuje rozhodující podíl VLP a s ním související oscilace některých výrazů, která se po roce 2008 zase mění.

Toto rozdělení na tři pomyslné úseky je výrazné zejména na obr. 6.3.7 a obr. 6.3.9. Toto rozdělení zároveň odpovídá obr. 4.5.4 na straně 58, a je tedy důsledkem složení zdrojových dat. Přitom více či méně ovlivňuje i řadu dalších, velice frekventovaných slov z jádra slovní zásoby, jak je patrné na frekvenčních průbězích lemmat *jestli* (obr. 6.5.16 na straně 161) a *třeba* (obr. 6.5.13 na straně 160), nebo dokonce na frekvenčním průběhu zájmen (obr. 6.5.5 na straně 156) a sloves (obr. 6.5.4 na straně 155). Tento fakt ukazuje možné dopady nevhodného složení korpusů, které se projevují zvláště v případě použité iterativní metody; ta preferuje výrazy s velkou oscilací, a poukazuje tak i na důsledky

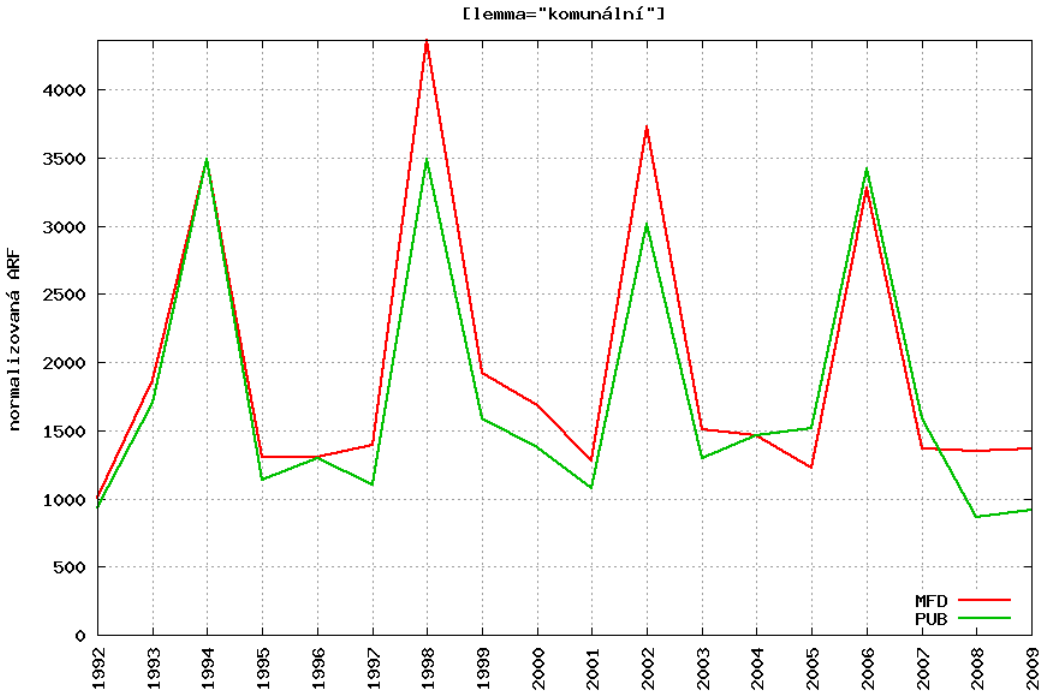
chyb a opomenutí ve zpracování korpusových dat. Praktickým výstupem této části jsou proto především návrhy několika opatření pro budoucí zpracování a kategorizaci textů v korpusech řady SYN, které shrneme dále v podkapitole 6.6.



Obrázek 6.3.1: Průběh normalizované ARF plurálu lemmatu *volba*.

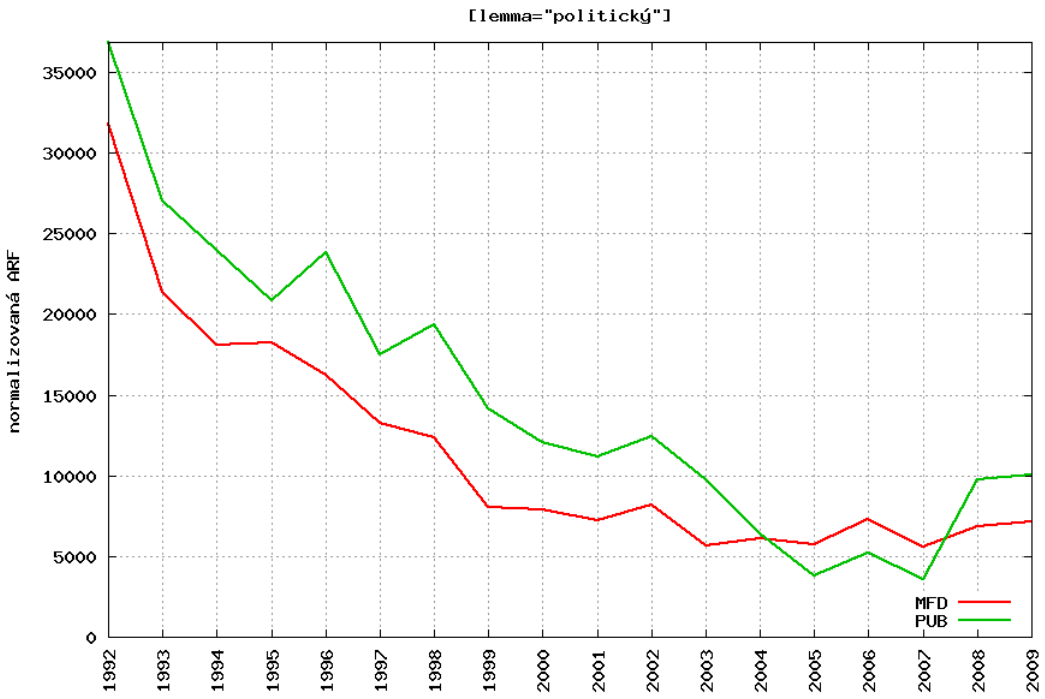


Obrázek 6.3.2: Průběh normalizované ARF singuláru lemmatu *volba*.

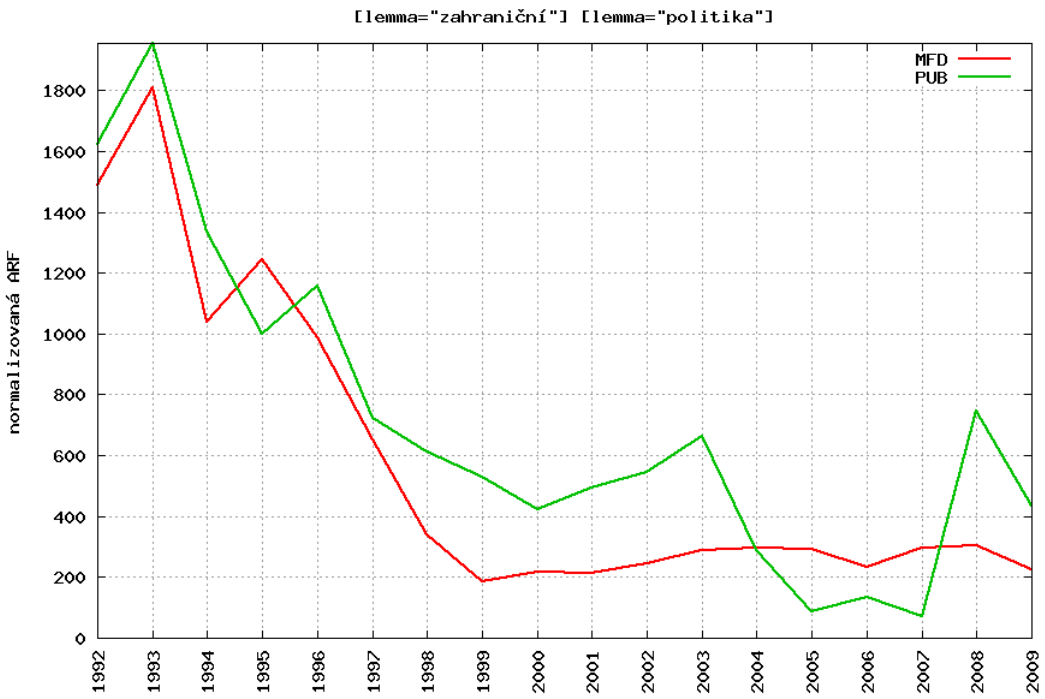


Obrázek 6.3.3: Průběh normalizované ARF lemmatu *komunální*.

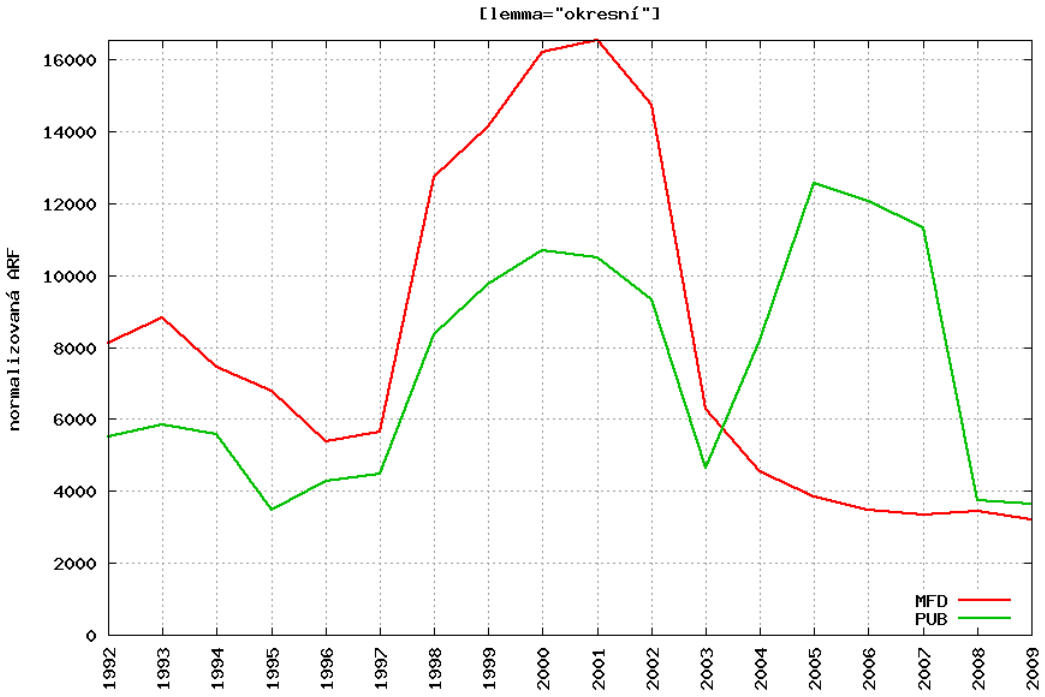
Obrázek 6.3.4: Průběh normalizované ARF lemmatu *lídr*.



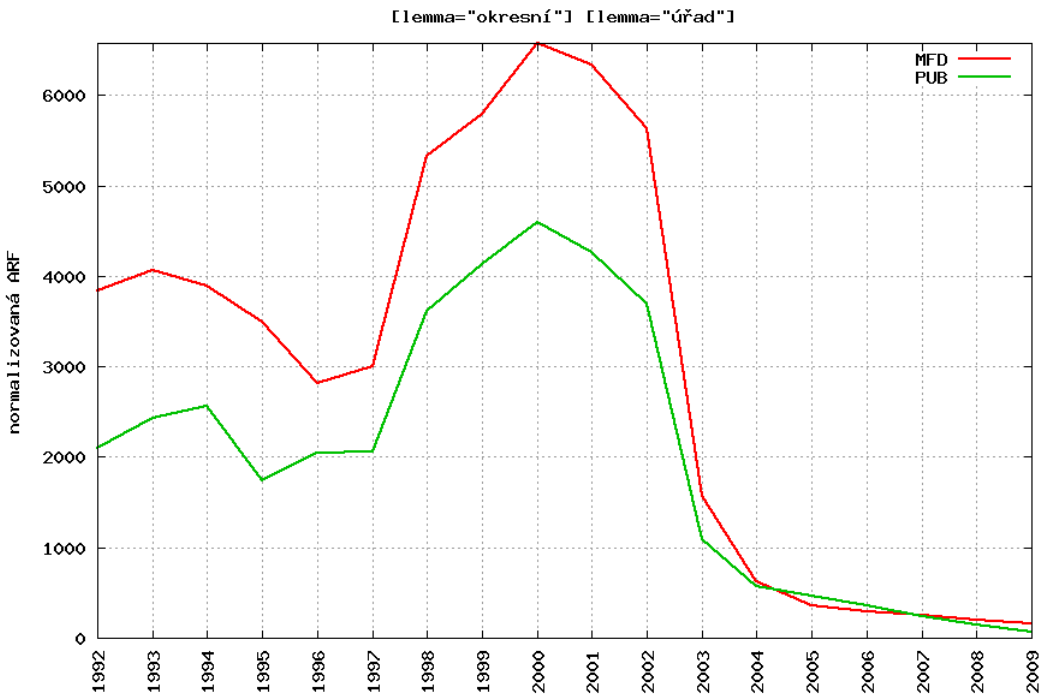
Obrázek 6.3.5: Průběh normalizované ARF lemmatu *politický*.



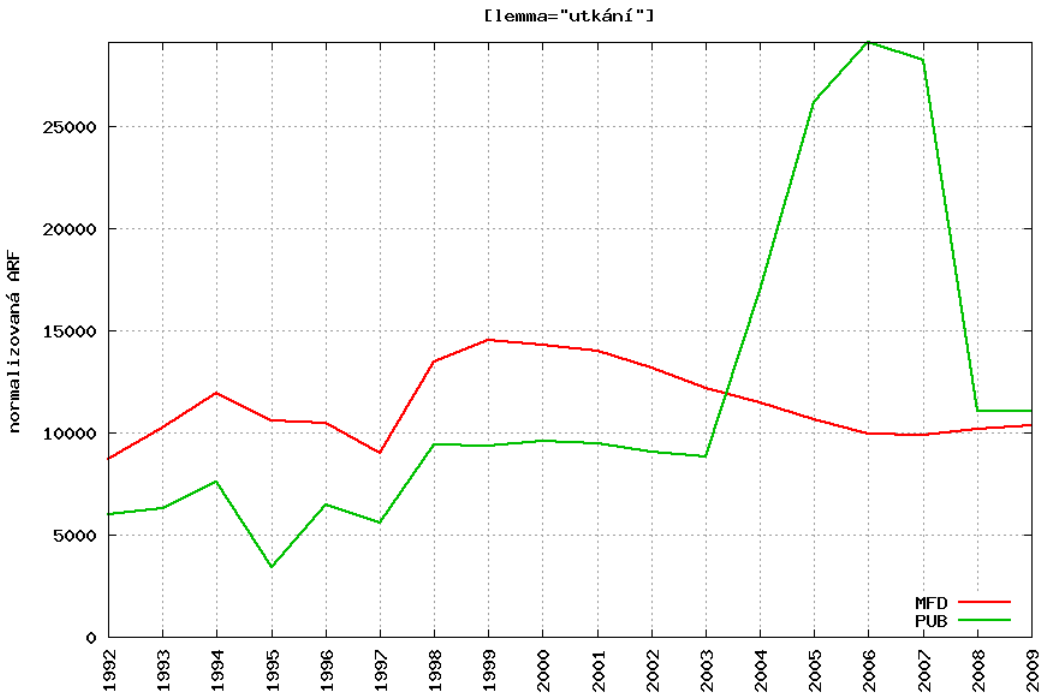
Obrázek 6.3.6: Průběh normalizované ARF kombinace *zahraniční politika*.



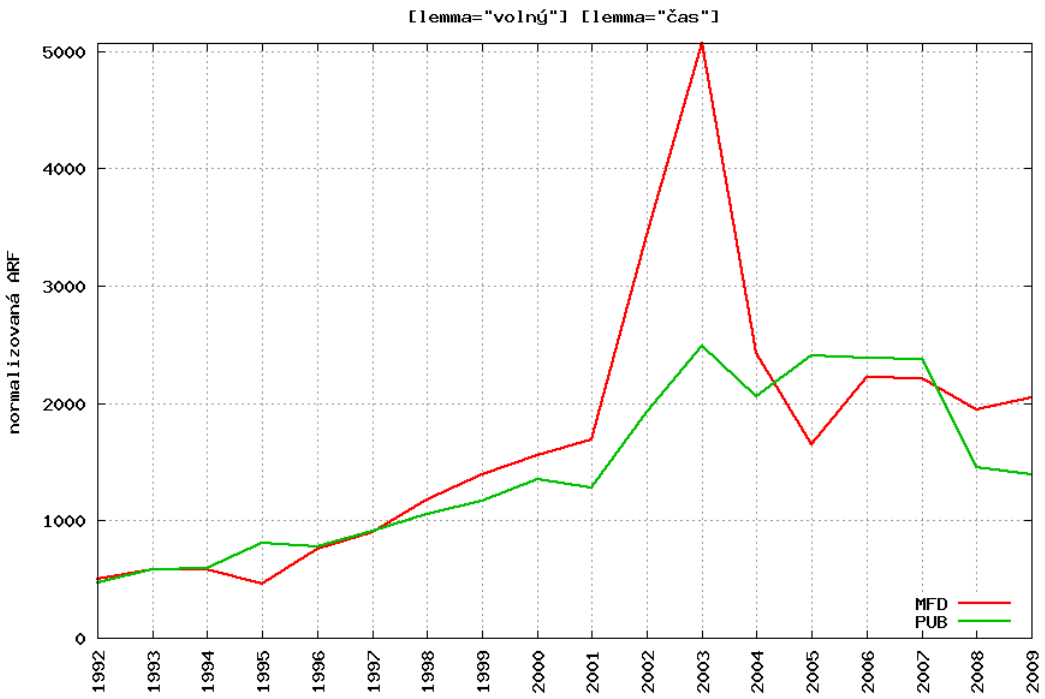
Obrázek 6.3.7: Průběh normalizované ARF lemmatu *okresní*.



Obrázek 6.3.8: Průběh normalizované ARF kombinace *okresní úřad*.



Obrázek 6.3.9: Průběh normalizované ARF lemmatu *utkání*.



Obrázek 6.3.10: Průběh normalizované ARF kombinace *volný čas*.

6.4 *tau* na publicistických subkorpusech

Tato podkapitola se zabývá rozborem výsledků metody *tau* aplikované na publicistické subkorpusy řady pub_RRRR a mf_RRRR. Pro výběr lemmat a kombinací pomocí *tau* je podstatná pravidelnost nárůstu nebo poklesu, ne jeho absolutní hodnota (viz oddíl 5.3.2). To umožňuje sledované vývojové tendence jasně identifikovat jako rostoucí či klesající, ve 3. sloupci výsledných tabulek je proto uvedeno znaménko + nebo − označující nárůst nebo pokles. Upozorněme také na fakt, že vzhledem k rozdílné povaze *tau* a iterativních *cbf*, *chi*, *ll* nenajdeme ve výsledných tabulkách v této podkapitole žádné z lemmat ani lexikálních kombinací z výsledných tabulek podkapitoly 6.3.

V tabulkách najdeme řadu výrazů s nulovou frekvencí v některém z pozorovaných let; jejich frekvence tedy buď roste od nuly v roce 1992, nebo naopak klesá k nule v roce 2009. Nejvyšší hodnota frekvence těchto výrazů dosažená na konci, resp. začátku tohoto období přitom není pro *tau* podstatná, a proto se tyto hodnoty mohou při stejném ohodnocení pomocí *tau* lišit i o několik řádů.

Při srovnání výsledných tabulek si můžeme všimnout také toho, že výrazy v tabulkách založených na subkorpusech řady pub_RRRR jsou obecně méně frekventované než výrazy v tabulkách založených na subkorpusech řady mf_RRRR, a to na úrovni lemmat i lexikálních kombinací. Příčinou je zřejmě proměnlivější složení subkorpusů řady pub_RRRR, které způsobuje větší náhodnost výsledků, a tedy odfiltrování frekventovanějších výrazů, kterých je ve srovnání s výrazy méně frekventovanými vždy výrazně méně. Stačí pak i velice malá ARF některých lemmat a její pravidelný nárůst ([P]*nákop*,[7] [P]*specialitka*) nebo pokles ([P]*hypertrofie*, [P]*obviňující*), který však při takto nízkých frekvencích může být dílem náhody. Totéž platí samozřejmě také na úrovni lexikálních kombinací, výmluvným příkladem je konkrétní věk u kombinace [P]*šestadvacetiletý mladík*. V těchto případech se také stává, že proti pravidelnému průběhu v jedné řadě subkorpusů stojí průběh dosti nepravidelný a s výraznou oscilací, což je další faktor snižující výpovědní hodnotu zjištěných frekvenčních posunů u málo frekventovaných slov.

Jistá míra nahodilosti v tom, který výraz bude vykazovat „nejpravidelnější“ nárůst nebo pokles, se projevuje také v malé shodě mezi výsledky založenými na subkorpusech řady pub_RRRR a mf_RRRR. Mezi prvními 50 tak nacházíme pouhá čtyři shodná lemmata ([MP]*developerský*, [MP]*hlavně*, [MP]*plánovat*, [MP]*webový*) a tři shodné kombinace ([MP]*nákupní centrum*, [MP]*webový stránka*, [MP]*zlínský kraj*). U řady výrazů se navíc opět projevuje nevhodné složení publicistických subkorpusů, jehož vliv na výsledky byl demonstrován v podkapitole 6.3. Nejvýraznější je vliv VLP v letech 2005–2007, což je vidět zejména na frekvenčním průběhu lemmatu *médium* na obr. 6.4.12, ale v menší míře i na průbězích některých velmi frekventovaných lemmat, například [M]*přidat*, [M]*umět*

[7] Dlouhé nakopnutí míče.

rank	lemma		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	hlavně	+	9664	10557	11105	11176	11189	12739	13661	13618	14380	15047	15415	17065	17817	19602	20566	20851	20773	21719
2	nastavení	+	43	40	116	153	199	227	273	319	344	424	398	473	481	555	651	681	759	861
3	plánovat	+	1897	2312	2540	3358	3644	3368	4038	4355	4657	4904	5040	5314	5809	6244	6236	6631	6903	6915
4	obviňující	–	54	64	37	35	34	32	32	19	17	16	14	14	8	4	1	0	0	0
5	hodně	+	42641	47256	52994	51241	53200	53869	54604	53822	55679	56698	59080	61482	61654	64661	67805	67967	71270	72110
6	rozjíždět	+	179	208	224	198	238	278	294	306	379	398	443	461	458	498	541	551	620	659
7	developerský	+	0	0	0	20	13	17	21	25	30	39	69	118	118	131	152	245	480	621
8	lei	–	108	75	47	40	34	33	24	15	15	8	13	12	8	7	6	6	0	0
9	webový	+	0	0	0	0	29	52	155	372	835	1054	1047	1282	1316	1488	1873	2520	2794	2773
10	nalákat	+	70	127	84	104	152	165	207	217	259	298	320	382	411	441	428	438	480	481
11	ustát	+	22	52	51	40	118	165	137	183	212	239	320	338	373	396	421	484	594	595
12	vláček	+	60	69	65	74	89	106	118	138	186	177	192	199	225	247	259	299	266	279
13	specialitka	+	0	0	0	0	0	6	8	6	10	12	13	14	15	17	21	24	25	25
14	multifunkční	+	0	0	14	25	18	59	44	73	99	116	178	240	286	372	512	544	417	557
15	přidružení	–	304	387	196	306	105	80	59	52	56	40	37	20	14	12	6	5	0	0
16	příslušet	–	813	809	741	602	594	468	433	410	364	346	346	344	277	253	218	169	202	228
17	gólmanský	+	0	0	5	10	16	19	20	16	26	30	36	34	60	54	72	74	89	101
18	posttotalitní	–	152	127	75	49	34	26	25	19	17	10	13	11	6	2	2	3	0	0
19	adrenalinový	+	0	0	0	0	5	7	8	22	54	91	154	228	265	337	314	405	329	469
20	skiareál	+	0	0	0	0	5	6	11	10	18	23	20	38	90	129	182	181	190	266
21	skousnout	+	0	0	0	5	0	6	6	10	15	18	34	46	35	44	69	80	114	165
22	multižánrový	+	0	0	0	0	0	7	5	7	8	18	21	39	34	43	56	65	76	89
23	kruh	–	4510	4157	3807	3798	3390	2956	2633	2505	2303	2284	2010	2425	1993	1774	1722	1549	1719	1672
24	nabírat	+	233	208	261	351	332	384	419	477	483	465	503	554	550	640	668	688	771	811
25	nadýchat	+	27	35	28	59	73	65	75	85	101	113	111	188	216	301	420	544	379	570

Tabulka 6.4.1: Úroveň lexikální, *pub_RRRR*, *tau*, 1. část.

rank	lemma		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	opilost	+	114	121	149	123	147	186	237	227	250	326	322	371	386	383	401	405	468	469
27	pomoci	+	9789	10533	11207	10746	10741	12217	13139	13875	13856	14238	15570	15255	15665	16513	16826	16895	17359	17477
28	pořídit	+	1138	1347	1747	1689	1864	2208	1994	2125	2132	2263	2466	2614	2899	2906	3038	3109	3325	3255
29	rozporný	–	390	364	303	286	217	214	167	162	144	147	106	86	53	25	22	17	13	51
30	úloha	–	3501	2885	2763	3240	2662	2245	1939	1782	1619	1596	1483	1392	1264	1133	1106	1101	1188	1000
31	výbor	–	14396	11828	10215	8805	9089	7321	7115	6313	6015	5925	5337	4812	5036	4805	4340	4318	4033	4597
32	vyslovený	–	455	358	266	262	238	199	183	156	146	140	111	113	104	91	97	103	63	89
33	záležitost	–	8976	9100	8943	7615	7568	7196	7487	7139	6901	6670	6092	5759	5807	5935	5567	5538	5121	4825
34	zařídit	+	932	1208	1212	1220	1233	1305	1387	1371	1424	1464	1559	1600	1632	1751	1870	1811	1707	2064
35	zateplení	+	27	29	42	44	58	97	122	100	106	113	134	140	212	306	370	410	341	709
36	zvládat	+	531	509	536	642	628	806	844	932	985	1036	1125	1161	1284	1386	1440	1355	1492	1469
37	kodifikovat	–	87	133	47	64	42	28	22	25	19	19	15	13	6	5	6	5	0	0
38	proíránský	–	222	191	107	59	107	72	30	16	12	8	6	5	2	0	2	0	0	0
39	zarezervovat	+	0	0	0	5	5	6	7	5	7	14	11	15	28	28	44	55	63	76
40	cz	+	0	0	5	15	81	267	606	1414	2986	4708	4945	5445	5604	4942	5438	8842	10696	11081
41	eventuální	–	829	734	657	657	531	436	333	272	223	250	213	178	164	161	158	148	177	114
42	nabrat	+	304	382	345	440	487	657	642	658	714	769	721	732	732	808	904	905	974	1026
43	navigace	+	38	40	37	59	50	63	67	72	93	93	96	115	102	126	182	320	468	342
44	odmala	+	11	23	14	25	16	22	34	36	42	66	66	77	79	82	87	117	139	165
45	pramenící	–	152	139	121	123	136	113	106	98	93	84	74	61	59	53	52	45	51	51
46	samosprávný	–	851	798	829	469	539	323	263	231	208	179	172	156	141	90	89	61	76	89
47	prokomunistický	–	119	98	42	44	29	15	16	14	9	6	8	8	4	2	2	0	0	0
48	hypertrofie	–	16	12	19	15	10	9	8	8	5	4	3	3	2	2	0	0	0	0
49	nákop	+	0	0	0	0	3	0	3	6	8	10	10	14	22	52	49	61	63	63
50	byro	–	184	87	65	54	45	26	15	16	23	13	12	11	8	1	1	2	0	0

Tabulka 6.4.2: Úroveň lexikální, *pub_RRRR*, tau, 2. část.

rank	lemma		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	chystat	+	4284	4563	5691	5328	6464	6647	7106	7380	8186	8525	8741	10308	11034	11247	11735	12068	12459	12561
2	jestli	+	5061	5249	5938	6594	7608	8546	8371	8565	8861	9342	9404	9948	10905	12315	13103	13397	14693	15645
3	webový	+	0	0	0	0	57	58	112	280	633	790	899	1106	1303	1478	1958	2329	2724	2822
4	čekat	+	17308	16360	17685	18996	19062	22908	25338	25434	26499	28126	29344	30726	32184	32855	33129	33244	34196	33525
5	sdělovací	–	2798	2504	2362	2445	1748	1269	853	621	454	440	404	280	228	232	203	189	148	132
6	týž	–	3610	3172	2641	2664	2606	2337	2115	1992	1844	1693	1566	1435	1370	1332	1286	1200	1208	1162
7	vyzkoušet	+	1002	1039	1256	1747	1824	2314	2477	2871	3008	2972	3115	3535	3745	4013	4092	4096	4233	4217
8	zahraniční	–	25478	20682	17814	20262	15007	13133	9036	8524	8367	8170	7614	7455	7553	7211	6800	6525	6362	6090
9	ustavení	–	1054	946	612	568	502	300	194	170	170	129	109	74	66	54	67	48	48	45
10	ale	+	146391	128785	142360	157165	168729	176072	199393	200345	199501	207181	213048	219320	229121	245405	251706	253641	260936	263500
11	hlavně	+	9172	9831	11532	12489	12725	14010	15839	15174	15642	16517	16697	18689	20157	21682	22201	23805	24927	24655
12	chtít	+	62391	63530	70665	75766	76565	77136	80183	77777	78815	81949	86452	89728	94150	98718	97777	101445	104105	104838
13	komunikovat	+	415	482	548	524	686	883	819	971	917	1041	1106	1159	1202	1232	1246	1355	1380	1480
14	li	–	15961	14988	13132	14673	10780	11171	7864	6995	6698	6407	6144	6003	5424	5201	5074	5191	4783	4780
15	ochutnat	+	86	93	107	87	229	265	316	348	369	353	419	580	730	773	875	882	986	1012
16	plánovat	+	2056	2671	2577	3101	3617	3456	4770	5375	5554	5866	5928	6325	7191	7631	7514	8155	8846	9073
17	popisovat	+	777	705	859	1004	1697	2498	3713	3807	3927	4256	4141	4658	5273	6444	6958	7281	7243	7835
18	sednout	+	328	538	709	568	788	941	1103	1137	1199	1236	1251	1374	1475	1686	1697	1787	1931	1736
19	trošku	+	225	297	258	306	464	571	731	808	980	1083	1052	1121	1297	1506	1872	1795	1981	2083
20	umět	+	4267	4285	4757	5677	5536	6330	6360	6748	6407	6799	6989	7378	8079	8792	8638	8939	9090	9134
21	začít	+	32094	34575	36079	39695	38620	39716	41214	42254	44990	46401	46032	49455	50526	52494	53450	52937	54491	57100
22	západoevropský	–	1866	2022	1718	1703	1360	969	526	417	383	338	284	251	208	240	210	205	202	168
23	mimoúrovňový	+	0	0	11	44	44	35	56	85	110	112	126	132	172	196	176	199	220	237
24	server	+	0	0	0	0	76	81	125	195	559	660	704	904	1048	995	1190	1700	2147	2047
25	bonusový	+	0	0	0	0	0	17	11	20	25	31	38	40	51	81	82	84	90	116

Tabulka 6.4.3: Úroveň lexikální, *mf_RRRR*, *tau*, 1. část.

rank	lemma		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	cyklostezka	+	0	0	0	0	0	17	189	274	378	496	653	853	1050	1217	1442	1578	1576	1825
27	smutnit	+	0	0	0	0	19	29	21	79	85	85	86	95	133	162	193	233	250	242
28	developerský	+	0	0	0	0	0	6	16	16	20	27	58	121	153	240	411	587	655	591
29	existence	–	4888	4062	3855	3843	3566	3156	3357	3082	3028	2815	2610	2525	2254	2263	2295	2217	2073	1940
30	ideální	+	1486	1688	1707	2533	2492	2752	2721	2833	2862	2852	3095	3333	3543	3709	3995	3989	4096	4105
31	najít	+	14751	13578	14335	15546	16812	19001	19730	20358	20269	20746	21813	23307	23830	25605	25898	26337	26177	26637
32	odehrát	+	2038	1892	2394	2576	3102	3151	3705	4116	4800	5754	5301	5540	5556	5901	5953	6322	6725	6751
33	pár	+	8257	6641	6829	8079	8339	9890	10075	10478	11296	11275	11751	12933	13939	14988	15269	16116	16344	16677
34	popsat	+	864	686	762	873	1875	3145	4382	3791	4270	4410	4706	5221	5892	7555	7996	8567	8912	9909
35	projít	+	3714	3691	4155	4105	5326	5297	5440	5765	6210	6347	6621	6846	7248	7668	7753	8099	8453	8377
36	představitel	–	19916	17918	15022	16332	12661	9971	7388	7080	6912	6571	5894	4707	4288	4114	3991	3443	3528	3615
37	přibýt	+	2073	1799	2319	2445	2841	2695	2876	3271	3119	3151	3309	3612	4228	4593	4860	5097	5340	5537
38	přidat	+	2297	3135	3028	4061	4087	4835	4858	4922	5234	5241	5340	5264	5563	5685	5659	6230	6516	6397
39	republika	–	46344	53328	44765	36333	28037	24956	23661	23462	22726	23372	23069	22932	21419	20390	19858	19693	18996	18207
40	tajemník	–	5510	4582	3866	4105	3693	2770	2729	2674	2486	2506	2370	2098	1814	1613	1592	1320	1374	1373
41	tam	+	18862	16880	16804	18254	20257	23075	27081	27476	28818	30872	32644	35092	37215	40880	42084	43668	43941	46231
42	tisk	–	7168	6344	5691	5197	4290	3785	3144	3051	3218	2796	2584	2147	1859	1878	1597	1706	1590	1335
43	už	+	90426	83934	103565	106596	105848	109639	118474	115100	116248	121684	132092	140755	145164	158148	162784	166995	170996	175298
44	vědět	+	24891	21999	25040	27293	29613	32440	33802	32873	33848	34487	33849	35523	36017	37651	37189	37702	38528	39706
45	venkovní	+	155	130	161	306	591	312	505	680	730	805	984	1015	1192	1318	1334	1481	1446	1687
46	vysvětlovat	+	3524	3951	5133	6376	6680	7282	9679	10241	10397	10435	10163	10612	11568	12174	12480	12721	13201	12894
47	zastávka	+	812	1354	891	1004	1614	1783	1876	1940	2119	2097	2366	2699	3257	3414	3452	3543	3516	3557
48	zeleň	+	328	334	408	437	769	802	1021	918	1110	1257	1316	1414	1701	1655	2583	2267	2580	2652
49	zvládnout	+	3075	2764	2953	3406	3528	4351	5111	5420	5720	5965	6386	6565	6495	6955	6947	7494	7610	7741
50	motorkář	+	0	0	0	44	44	98	117	123	157	199	190	255	247	311	428	523	723	614

Tabulka 6.4.4: Úroveň lexikální, *mf_RRRR*, *tau*, 2. část.

rank	kombinace		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	hromadný sdělovací	–	190	81	75	64	31	28	26	20	18	14	15	7	7	5	5	5	0	0
2	webový stránka	+	0	0	0	0	24	43	114	290	654	809	886	1105	1142	1290	1621	2198	2364	2280
3	liberalizace cena	–	379	145	130	123	99	80	27	17	19	10	7	4	2	1	2	0	0	0
4	evropský integrační	–	65	64	65	59	37	30	20	21	20	18	18	17	8	4	4	2	0	0
5	nástupnický stát	–	596	370	112	89	42	41	26	22	22	15	16	13	6	7	2	5	0	0
6	privatizovaný podnik	–	255	237	144	163	126	67	45	24	17	21	14	10	6	6	1	2	0	0
7	proces privatizace	–	249	254	121	237	120	71	83	45	45	31	20	15	8	7	5	3	0	0
8	auto vyprostit	+	0	0	0	0	0	2	2	4	6	6	9	11	14	18	22	22	25	25
9	dosud existující	–	92	92	51	64	34	33	31	23	25	18	19	16	13	7	5	5	0	0
10	silný opilost	+	0	0	0	0	5	7	7	7	8	9	14	13	18	20	20	24	38	38
11	privatizační metoda	–	70	58	47	44	37	15	6	7	4	4	3	2	1	0	0	0	0	0
12	nákupní centrum	+	16	40	47	44	86	87	97	118	182	227	277	350	347	359	410	518	695	469
13	rozpad federace	–	656	254	196	148	118	100	53	47	26	29	32	25	11	12	10	7	0	0
14	daňový soustava	–	537	312	168	128	92	87	59	48	54	46	49	21	12	11	4	8	0	0
15	vnitřní rozpor	–	87	98	56	74	52	45	32	33	23	21	22	20	11	6	5	3	0	0
16	nacionalistický síla	–	33	29	28	25	13	7	5	6	5	5	4	2	1	0	0	0	0	0
17	trojitý skok	–	38	64	37	59	29	20	20	18	15	9	9	8	8	8	7	6	0	0
18	letní posila	+	0	0	0	0	3	4	11	16	19	41	42	49	53	82	60	110	89	101
19	přežít střet	+	0	0	0	0	0	4	4	5	6	6	11	7	13	13	14	20	38	38
20	ediční činnost	–	38	29	28	25	18	15	14	13	13	11	12	7	10	4	4	5	0	0
21	právně postižitelný	–	16	23	14	10	10	9	8	6	6	3	4	2	1	2	1	1	0	0
22	už odmala	+	0	0	0	0	3	4	5	7	9	6	10	10	11	13	17	23	51	25
23	odpočinkový zóna	+	0	0	0	0	8	9	13	8	14	20	26	24	50	65	76	84	89	101
24	šestadvacetiletý mladík	+	0	0	0	0	5	6	8	10	11	12	14	15	17	22	30	24	25	25
25	hip hop	+	0	6	9	10	13	20	27	38	54	70	85	111	149	150	163	204	215	114

Tabulka 6.4.5: Úroveň lexikálních kombinací, *pub_RRRR*, `tau`, 1. část.

rank	kombinace		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	orgán státní	–	575	451	387	681	382	290	237	221	215	209	178	137	130	142	117	116	114	89
27	podle sdělení	–	1014	746	704	519	369	288	257	272	232	155	145	102	132	143	94	91	51	76
28	zásadně odlišný	–	54	23	23	30	26	22	18	16	16	15	13	9	6	6	5	3	0	0
29	vyjádřit přesvědčení	–	369	266	303	237	257	186	150	126	131	108	93	83	45	38	32	24	38	13
30	privatizační fond	–	650	989	629	484	217	108	51	50	45	41	18	15	10	4	2	0	13	0
31	vyřešení otázka	–	54	46	51	44	37	24	18	15	21	14	11	11	8	8	7	8	0	0
32	o přidružení	–	222	231	93	148	42	33	21	17	18	18	11	9	5	2	2	2	0	0
33	demokratický ústava	–	81	52	23	30	10	20	9	8	7	5	3	4	4	1	1	0	0	0
34	developerský firma	+	0	0	0	5	5	7	11	6	8	11	30	49	41	64	76	84	152	165
35	pluralitní demokracie	–	98	116	84	40	24	24	26	20	11	9	8	5	7	5	5	2	0	0
36	prezidentův pravomoc	–	27	23	28	20	13	13	11	9	8	4	5	4	2	2	2	2	0	0
37	počasí přilákat	+	0	0	0	0	5	7	9	10	12	17	15	20	23	38	35	36	38	38
38	podvodník okrást	+	0	0	0	0	0	0	3	5	4	6	8	10	10	12	15	21	25	25
39	vážný nebezpečí	–	92	92	89	74	84	65	61	59	47	52	57	44	37	39	35	29	25	25
40	čtvrtfinálový série	+	0	0	9	5	8	11	15	26	26	51	43	46	62	84	102	109	139	114
41	politický aspekt	–	76	58	65	54	37	35	19	21	11	17	8	5	6	4	5	3	0	0
42	zákonodárný orgán	–	152	69	98	79	55	37	24	18	11	16	14	9	8	8	2	3	0	0
43	celní unie	–	667	283	340	904	304	230	133	79	51	19	17	12	11	8	2	2	0	0
44	územní samospráva	–	103	69	89	54	68	35	32	23	28	23	26	21	20	19	16	5	0	0
45	demokratický instituce	–	206	145	89	94	89	84	57	67	54	37	31	43	15	5	7	2	0	0
46	nadace partnerství	+	0	0	0	0	5	13	18	27	29	31	43	36	65	106	145	189	126	152
47	prostřednictvím internetový	+	0	0	0	0	8	11	23	29	40	31	37	37	44	44	44	55	76	89
48	zlínský kraj	+	0	0	0	0	0	6	18	32	309	461	641	679	756	848	884	1462	999	937
49	všeobecný dohoda	–	260	358	214	138	31	20	15	13	8	8	5	5	1	0	5	0	0	0
50	hasič zachraňovat	+	0	0	5	0	5	6	5	7	7	8	18	20	20	20	22	33	25	38

Tabulka 6.4.6: Úroveň lexikálních kombinací, *pub_RRRR*, `tau`, 2. část.

rank	kombinace		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	webový stránka	+	0	0	0	0	44	58	85	236	508	618	804	968	1147	1318	1697	2026	2404	2448
2	privatizace státní	–	345	260	225	175	153	98	85	70	49	41	39	28	30	24	23	20	22	13
3	bytový dům	+	17	74	86	87	127	202	239	296	369	487	729	854	927	1036	948	1051	1090	1053
4	sdělovací prostředek	–	2695	2448	2169	2271	1678	1264	805	603	436	419	370	257	210	212	174	165	128	121
5	vysvětlit mluvčí	+	0	0	32	0	83	110	202	247	298	297	396	437	445	529	554	581	609	762
6	evropský fond	+	0	0	0	0	6	0	13	45	83	116	139	323	524	668	843	981	1226	1400
7	popsat mluvčí	+	0	0	0	0	0	12	19	29	34	41	98	146	161	246	365	533	517	770
8	mimoúrovňový křižovatka	+	0	0	11	0	32	23	35	52	65	68	73	81	110	145	136	153	170	184
9	lezecký stěna	+	0	0	0	0	0	6	11	11	16	19	32	51	62	93	63	121	126	141
10	zlínský kraj	+	0	0	0	0	0	6	56	87	633	604	697	771	835	1028	1028	1089	1188	1250
11	liberecký kraj	+	0	0	0	0	0	6	13	13	300	848	888	1062	1127	1339	1330	1344	1350	1391
12	národní majetek	–	3178	3580	3790	2838	2549	1506	1135	765	631	539	421	312	167	170	94	68	56	45
13	tisíc marka	–	587	742	698	480	458	433	322	368	311	270	71	45	30	18	17	16	8	11
14	celý záležitost	–	1676	1948	1235	1572	1182	1004	702	700	559	664	513	463	434	392	342	332	262	329
15	nákupní centrum	+	35	74	43	87	64	92	98	141	237	295	340	450	591	757	741	814	839	889
16	milión marka	–	1503	1410	1396	1572	1322	992	470	419	367	359	54	36	17	12	10	10	6	9
17	zahraniční kapitál	–	1503	1113	945	961	521	329	266	231	114	91	53	72	47	44	31	38	38	16
18	výměna okno	+	0	0	0	0	19	12	40	27	49	77	75	102	155	188	248	265	385	482
19	rušný silnice	+	0	0	0	0	6	12	29	38	31	41	49	57	82	87	88	109	114	92
20	čerpání dotace	+	0	0	0	0	0	0	3	11	13	17	19	32	56	59	82	82	82	90
21	ředitelství silnice	+	0	0	0	175	19	138	353	406	436	678	610	771	713	880	1013	1059	1146	1160
22	docela dost	+	0	0	0	0	19	35	35	54	65	60	96	106	142	141	147	155	214	197
23	volnočasový aktivita	+	0	0	0	0	0	0	5	18	29	46	73	100	105	107	136	193	206	202
24	spolkový republika	–	553	371	301	306	254	294	167	162	161	137	111	76	105	101	82	72	66	63
25	zahraniční firma	–	1244	1002	1020	611	788	462	439	431	376	351	259	304	281	275	258	213	154	130

Tabulka 6.4.7: Úroveň lexikálních kombinací, *mf_RRRR*, `tau`, 1. část.

rank	kombinace		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	nový cyklostezka	+	0	0	0	0	0	0	13	25	27	58	58	117	116	157	191	195	202	231
27	začít stavět	+	207	241	505	349	451	531	603	603	689	728	868	1008	971	1213	1292	1226	1284	1342
28	popřít zpráva	–	104	93	64	87	51	40	19	18	13	17	17	13	9	10	6	2	4	0
29	stránka www	+	0	0	0	0	0	17	13	135	269	523	777	671	743	949	1062	1495	1655	1812
30	evropský dotace	+	0	0	0	0	0	0	0	7	7	21	26	34	105	258	329	460	747	1106
31	uvidět jestli	+	0	0	43	0	83	40	101	110	112	129	150	117	168	184	197	219	234	300
32	slavit postup	+	0	0	11	0	13	23	29	43	49	46	53	55	54	71	67	84	72	92
33	bezpečnostní kamera	+	0	0	0	0	13	29	29	43	40	75	100	189	197	206	231	285	272	244
34	sběrný dvůr	+	0	0	0	0	13	58	120	119	103	129	128	174	174	206	227	289	377	419
35	dotační program	+	0	0	0	0	0	12	5	18	27	39	39	60	79	125	138	123	150	246
36	pokrýt dotace	+	0	0	0	0	0	0	11	9	13	17	26	38	49	79	86	80	138	166
37	fond národní	–	1399	2282	2556	2052	1913	1062	1013	660	573	500	387	280	167	176	90	68	66	40
38	konečný rozhodnutí	–	587	556	430	742	381	352	284	265	239	259	231	193	189	182	185	173	150	159
39	tento souvislost	–	3627	3228	2931	2926	1831	1368	811	615	725	658	569	429	440	456	395	344	351	327
40	doprovodný program	+	69	19	21	0	108	173	213	363	506	531	539	737	863	749	917	1125	1168	1183
41	kompletní rekonstrukce	+	0	0	21	0	44	29	56	67	72	133	145	231	210	238	243	231	260	302
42	knižní velkoobchod	–	207	352	236	87	70	46	50	34	31	19	11	6	4	4	6	4	2	0
43	odpočinkový místo	+	0	0	0	0	0	6	16	22	22	44	32	28	54	55	57	64	64	92
44	cyklostezka podél	+	0	0	0	0	0	0	5	4	9	10	11	15	24	36	36	74	56	85
45	mluvčí přerovský	+	0	0	0	0	0	12	13	13	13	19	13	64	82	95	115	173	134	204
46	nechat zpracovat	+	35	0	21	44	89	52	117	117	166	174	171	204	217	283	279	307	290	361
47	autobusový zastávka	+	52	93	64	131	159	167	213	211	253	263	321	338	445	430	445	408	477	493
48	úplně jiný	+	639	575	601	611	833	935	1114	1299	1110	1222	1305	1323	1366	1385	1385	1521	1665	1673
49	kvůli bezpečnost	+	0	0	54	87	89	63	112	110	114	118	141	149	155	119	176	187	190	202
50	logistický centrum	+	0	0	0	0	0	0	0	4	16	31	24	34	58	93	111	117	134	190

Tabulka 6.4.8: Úroveň lexikálních kombinací, *mf_RRRR*, `tau`, 2. část.

(obr. 6.4.6), [M]*vědět*, [M]*vyzkoušet* nebo [M]*zvládnout*. Znatelné je také popisované rozdělení frekvenčního průběhu některých výrazů na tři pomyslné úseky (obr. 6.4.1, obr. 6.4.2, obr. 6.4.5 nebo obr. 6.4.6).

Úvodem diskuse konkrétních výrazů upozorněme na lemma [P]*hodně*, pod nímž jsou v použité verzi lemmatizace zahrnuty také některé výskyty tvarů *víc/více* a všechny výskyty tvarů *nejvíc/nejvíce*. Tvary *nejvíc/nejvíce* jsou tedy považovány vždy za superlativ adverbia *hodně*, tvary *víc/více* jsou lemmatizovány podle kontextu buď jako komparativ adverbia *hodně* (nikdy však *mnoho* nebo *moc*), nebo jako samostatná číslovka. Protože je lemmatizace komparativu a superlativu sporná a náchylná k chybám, zaměřili jsme se na pozitiv, jehož frekvenční nárůst je ve srovnání s ostatními dvěma stupni výrazný, jak je vidět na obr. 6.4.1 a obr. 6.4.2.

Lemmata a kombinace nalezené ve výsledných tabulkách můžeme rozdělit do tří hlavních skupin:

1. běžná, frekventovaná lemmata;
2. výrazy vykazující alespoň v jedné řadě subkorpusů (pub_RRRR nebo mf_RRRR) pravidelný nárůst;
3. výrazy vykazující alespoň v jedné řadě subkorpusů (pub_RRRR nebo mf_RRRR) pravidelný pokles.

Jazykově nejzajímavější je pravidelný a často výrazný nárůst[8] lemmat 1. skupiny, mezi nimiž je řada sloves: [M]*ale* (obr. 6.4.3), [M]*čekat*, [M]*chtít*, [M]*chystat*, [M]*ideální*, [M]*jestli* (obr. 6.5.16 na straně 161), [P]*nabírat*, [P]*nabrat*, [M]*najít*, [P]*nalákat*, [M]*pár*, [MP]*plánovat*, [P]*pomoci*, [P]*pořídit*, [M]*projít* (obr. 6.4.4), [M]*přibýt*, [M]*přidat*, [P]*rozjíždět* (obr. 6.4.5), [M]*sednout*, [P]*skousnout*, [M]*tam*, [M]*umět* (obr. 6.4.6), [P]*ustát*, [M]*už* (obr. 6.2.2 na straně 96), [M]*vědět*, [M]*vyzkoušet*, [M]*začít* (obr. 6.4.7), [P]*zařídit*, [P]*zvládat*, [M]*zvládnout*.

Ještě více frekventovaných lemmat najdeme v tabulkách založených na *taumed* v podkapitole 6.5. Na tomto místě proto uvádíme pouze část diskuse a poznamenáváme, že jejich nárůst má několik příčin. Jednou z nich je nahrazování jednoho lemmatu jiným, jehož význam je stejný nebo podobný. To však většinou k vysvětlení nestačí, například frekvenční nárůst lemmatu [M]*už* je výrazně vyšší než pokles konkurenčního *již*.

Další příčinou frekvenčního nárůstu běžných slov je pronikání neformálního způsobu vyjadřování do publicistiky, což ilustrují výrazy [M]*docela dost*, [MP]*hlavně*, [P]*hodně*, [P]*specialitka*, [M]*trošku*, [M]*úplně jiný*, [M]*uvidět jestli*,[9] jejichž frekvenční nárůst za sledované období je často dvojnásobný nebo větší. Velice častý je výskyt těchto výrazů v přímé řeči a rozhovorech, nejvyšší relativní nárůst přitom vykazuje kombinace [M]*docela dost*,

[8] O pokles jde u lemmat 1. skupiny pouze v jediném případě, a tím je příklonka [M]*li*.

[9] Používá se typicky ve tvaru *uvidíme, jestli*, tedy s čárkou, která však byla při výpočtu kombinací ignorována.

která se v nich vyskytuje téměř výhradně. Pro rozhovory a osobní charakteristiky je typické také používání výrazů [P]*odmala*, [P]*už odmala*. Řada frekventovaných výrazů se však hojně vyskytuje i mimo přímou řeč, tedy v publicistických textech jako takových, přestože je vliv přímé řeči značný i u velice frekventovaných výrazů z jádra slovní zásoby, jako je například lemma [M]*chtít*, které se v ní vyskytuje téměř v polovině případů. Míru nárůstu množství přímé řeči v publicistice jsme se pokoušeli kvantifikovat, avšak neúspěšně; ukázalo se totiž, že se nelze opřít o formální kritéria, jakými jsou například uvozovky. Přesto však lze její podíl vysledovat nepřímo na frekvenčním nárůstu výrazů [M]*mluvčí přerovský*, [M]*popisovat*, [M]*popsat* (obr. 6.4.9), [M]*popsat mluvčí* (obr. 6.4.10), [M]*vysvětlit mluvčí*, [M]*vysvětlovat*, z nichž většina uzavírá přímou řeč ve spojeních „popsal(a) XY“ nebo „vysvětluje XY“.

Další příčinou je stále častější užívání řady sloves v přeneseném významu. Typicky jde o spojení významově oslabeného, kategoriálního slovesa se substantivem, s nímž tvoří sémantický predikát věty (Martínek, 2011). Například sloveso [P]*ustát* má v subkorpusu pub_1992 většinou doslovný význam „zůstat stát, nespadnout“, zatímco v subkorpusu pub_2009 již zřetelně převažuje význam přenesený týkající se politiky (*ustát aféru*), sportu (*ustát tlak*) i běžného života (*ustát chemoterapii*). Podobně sloveso [P]*nadýchat* je v subkorpusu pub_2009 ve většině výskytů spojeno s alkoholem (a nejde tedy o reflexivum), zatímco v subkorpusu pub_1992 není na toto spojení doklad ani jeden, typicky jde o nadýchání se vzduchu nebo plynu.

Tyto příklady naznačují, že sledovanému nárůstu významově oslabených užití těchto sloves napomáhá množství šablonovitých spojení, která jsou pro publicistiku (zvláště sportovní) typická: *chystat akci/opatření/změnu*, *nabrat sílu/směr/zpoždění*, *nalákat klienty/turisty/voliče*, *projít kontrolou/rekonstrukcí/zkouškou*, *rozjíždět investice/projekt/výrobu*, *skousnout porážku*, *zařídit gól/remízu/vítězství/výhru/vyrovnání*, *zvládat emoce/situaci/zápas* apod.

Martínek (2011, str. 267) také uvádí hypotézu, že slovesa v kategoriálním užití vznikají desémantizací plnovýznamových sloves a jejich ustalováním v konkrétních syntagmatech, zatímco kategoriální užití jiných sloves naopak pozvolna zaniká; obecněji jde o posun směrem ke gramatikalizaci plnovýznamových sloves. Na počátku desémantizace přitom může být snaha o novost, neotřelost nebo dodání expresivity, později však může kategoriální užití daného slovesa tyto vlastnosti ztratit a stát se tak běžnou součástí jazyka. Jako téma pro další studii se proto nabízí podrobnější sledování a kvantifikace procesu postupných změn lexikálního významu těchto sloves právě na publicistickém materiálu z let 1992–2009, který se pro tyto účely zdá být vhodný.

Velkou část 2. skupiny tvoří výrazy spojené s nástupem internetu, změnou životního stylu a využíváním volného času: [P]*adrenalinový*, [M]*bonusový*, [P]*cz*,[10] [M]*doprovodný program*, [P]*hip hop*, [M]*komunikovat*, [M]*lezecký stěna*, [M]*logistický centrum*, [P]*multifunkční*,

[10] Jde o součást webových adres, které byly v místě tečky tokenizací chybně rozděleny na několik částí.

P*multižánrový*, P*nadace partnerství*,[11] MP*nákupní centrum*, P*navigace*, M*odpočinkový místo*, P*odpočinkový zóna*, M*ochutnat*,[12] P*počasí přilákat*, P*prostřednictvím internetový*, M*sběrný dvůr*, M*server*, P*skiareál*, M*stránka www*, P*vláček*,[13] M*volnočasový aktivita*, MP*webový*, MP*webový stránka*, P*zarezervovat*, M*zeleň*. Jejich přítomnost ve výsledných tabulkách není žádným překvapením, výrazný nárůst zaznamenala také řada dalších výrazů (*celebrita*, *euro*, *mobil* atd.), ten však není tak pravidelný. Do této skupiny patří i adjektivum M*venkovní*, ačkoli není tematicky tak vyhraněné; pojí se nejen s „volnočasovými“ substantivy *areál*, *bazén*, *expozice*, *voliéra*, *výběh*, ale i obecným *teplota* nebo sportovním *zápas*.

Do 2. skupiny dále patří řada výrazů z oblasti sportu: P*čtvrtfinálový série*, P*gólmanský*, P*letní posila*, P*nákop*, P*nastavení*,[14] M*odehrát*, M*slavit postup*. Patří sem také sloveso M*smutnit*, protože v současné publicistice „smutní“ téměř výhradně sportovec nebo trenér po závodě či zápase.

Ve 2. skupině najdeme také několik výrazů z dopravy: M*mimoúrovňový*, M*mimoúrovňový křižovatka*, M*rušný silnice*, M*ředitelství silnice* (součást spojení *Ředitelství silnic a dálnic*). Další výrazy souvisejí kromě dopravy také s měnící se realitou a životním stylem: M*cyklostezka*, M*cyklostezka podél*, M*nový cyklostezka*, M*motorkář*, zatímco jiné mají spíše regionální charakter: M*autobusový zastávka*, M*zastávka*. S dopravou souvisejí také některé z výrazů s kriminální a bezpečnostní tematikou: P*auto vyprostit*, M*bezpečnostní kamera*, P*hasič zachraňovat*, M*kvůli bezpečnost*, P*nadýchat*, P*opilost*, P*podvodník okrást,* P*přežít střet*, P*silný opilost*, P*šestadvacetiletý mladík*.

Mezi častá témata současné publicistiky patří také čerpání dotací, což dokumentuje frekvenční nárůst výrazů M*čerpání dotace*, M*dotační program*, M*evropský dotace*, M*evropský fond*, M*pokrýt dotace*. Další skupina výrazů se týká stavebnictví: M*bytový dům*, MP*developerský*, P*developerský firma*, M*kompletní rekonstrukce*, M*výměna okno*, M*začít stavět*, P*zateplení*.

Pravidelný frekvenční nárůst se však téměř nedotýká výrazů z oblasti politiky. Jediným výrazem typickým pro jazyk publicistiky z této oblasti je kombinace M*nechat zpracovat*, která bývá doplněna substantivy *posudek*, *projekt*, *studie*. Kombinace M*liberecký kraj*, MP*zlínský kraj* jsou jiného typu, pravidelnost nárůstu je v jejich případě navíc usnadněna politickou realitou, tedy tím, že kraje v 90. letech ještě neexistovaly.

Výrazy 3. skupiny, které alespoň v jedné řadě subkorpusů zaznamenaly pravidelný pokles, můžeme podle příčin tohoto poklesu dále rozdělit na tři podskupiny. Toto rozdělení je pochopitelně pouze orientační, protože řada výrazů stojí na pomezí některých podskupin. Příkladem (byť netypickým) může být adjektivum P*proíránský*, které se na počátku 90. let vyskytuje v množství obměn poněkud šroubovaného

[11] Nadace Partnerství se zabývá podporou projektů na ochranu životního prostředí.
[12] Jiný než doslovný význam se vyskytuje pouze ojediněle.
[13] Tato zdrobnělina označuje zpravidla výletní vláček nebo parní vlak.
[14] Část výskytů se netýká prodloužení ve sportu, ale nastavení přístrojů, systému nebo zákona.

spojení „(příslušníci) proíránské strany/organizace/hnutí Hizballáh“. Protože má ale samotné lemma *Hizballáh* spíše periodický průběh odpovídající skutečným událostem, je pozorovaný frekvenční pokles adjektiva [P]*proíránský* způsoben především tím, že se v tomto spojení přestalo používat. Není však zřejmé, zda jde o snahu o zjednodušení vyjadřování (a tedy příčinu primárně jazykovou), o důsledek zvýšení povědomí o této organizaci (a tedy o odraz reality), nebo spíše o projev politické korektnosti, případně o kombinaci těchto faktorů.

V první podskupině jsou výrazy, jejichž klesající frekvence pouze odráží měnící se realitu: [P]*byro*, [P]*celní unie* (mezi ČR a SR), [P]*daňový soustava* (nová daňová soustava od 1. 1. 1993), [M]*fond národní* (součást názvu *Fond národního majetku*), [M]*knižní velkoobchod*, [P]*liberalizace cena*, [M]*milión marka*, [M]*národní majetek* (většinou jde o součást názvu *Fond národního majetku*), [P]*nástupnický stát*, [P]*o přidružení* (součást spojení *dohoda o přidružení* ČSFR/ČR k ES/EU), [P]*posttotalitní,* [M]*privatizace státní*, [P]*privatizační fond*, [P]*privatizační metoda*, [P]*privatizovaný podnik*, [P]*proces privatizace*, [P]*prokomunistický*, [P]*přidružení*, [P]*rozpad federace*, [P]*samosprávný* (v 1. pol. 90. let jde často o součást názvu politické strany HSD-SMS), [M]*tisíc marka*, [P]*všeobecný dohoda*; poslední jmenovaná kombinace je součástí názvu Všeobecná dohoda o clech a obchodu (GATT), která byla později nahrazena Světovou obchodní organizací (WTO).

Druhou podskupinu tvoří výrazy, které odrážejí tematické zaměření publicistiky zejména na počátku 90. let, diskuse o rodící se demokracii apod.: [P]*demokratický instituce*, [P]*demokratický ústava*, [P]*nacionalistický síla* (jde ale zároveň o jazyk, tento výraz se dnes již moc neužívá), [P]*pluralitní demokracie*, [P]*právně postižitelný* (vyskytuje se převážně ve spojení s negací), [P]*prezidentův pravomoc*, [M]*republika* (výskyty pocházejí často ze spojení *Česká republika*), [M]*tajemník*, [P]*územní samospráva*, [P]*výbor*, [M]*zahraniční*, [M]*zahraniční firma*, [M]*zahraniční kapitál*, [M]*západoevropský*. Do této podskupiny lze zařadit také kombinaci [P]*trojitý skok*, která je výjimkou z její zahraničněpolitické a ekonomické orientace. Jde také o jediný výraz z oblasti sportu (jde ovšem o krasobruslení), který je ve výsledných tabulkách označen znaménkem minus. Tento výraz, stejně jako například kombinace [P]*ediční činnost*, také zaznamenal silný výkyv v MFD ročníku 1995, který je opět způsoben jeho složením. Netypické je také lemma [P]*lei*, jehož pravidelně klesající frekvenční průběh považujeme v zásadě za náhodný.

Poslední podskupina 3. skupiny výrazů ukazuje větší formálnost jazyka publicistiky 90. let ve srovnání s publicistikou současnou, kterou jsme se zabývali výše. Jde zejména o výrazy [M]*celý záležitost*, [P]*dosud existující*, [P]*eventuální*, [M]*existence*, [P]*hypertrofie*,[15] [P]*kodifikovat*, [M]*konečný rozhodnutí*, [P]*kruh*,[16] [P]*obviňující*, [P]*podle sdělení* (protipól současného „popsal(a) mluvčí, vysvětlil(a) mluvčí“), [P]*politický aspekt,* [M]*popřít zpráva*,

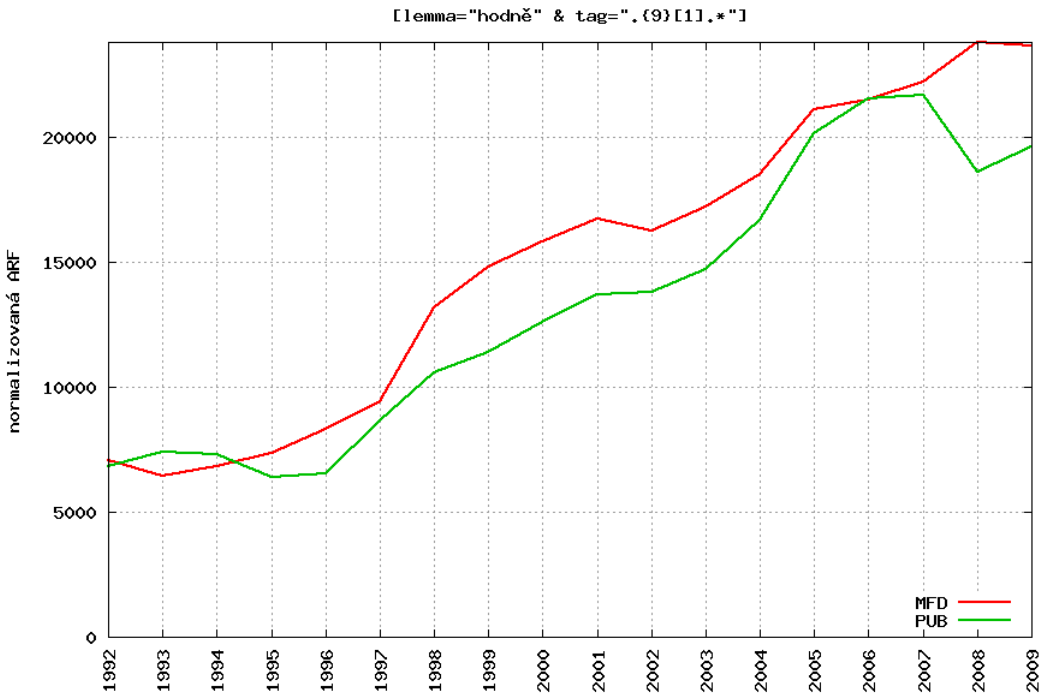
[15] Většina výskytů, zvláště z počátku 90. let, je užita v přeneseném, nikoli lékařském smyslu.

[16] K pozorovanému poklesu došlo v plurálu, týká se politických, vládních a jiných kruhů; v singuláru frekvence osciluje, ale neklesá.

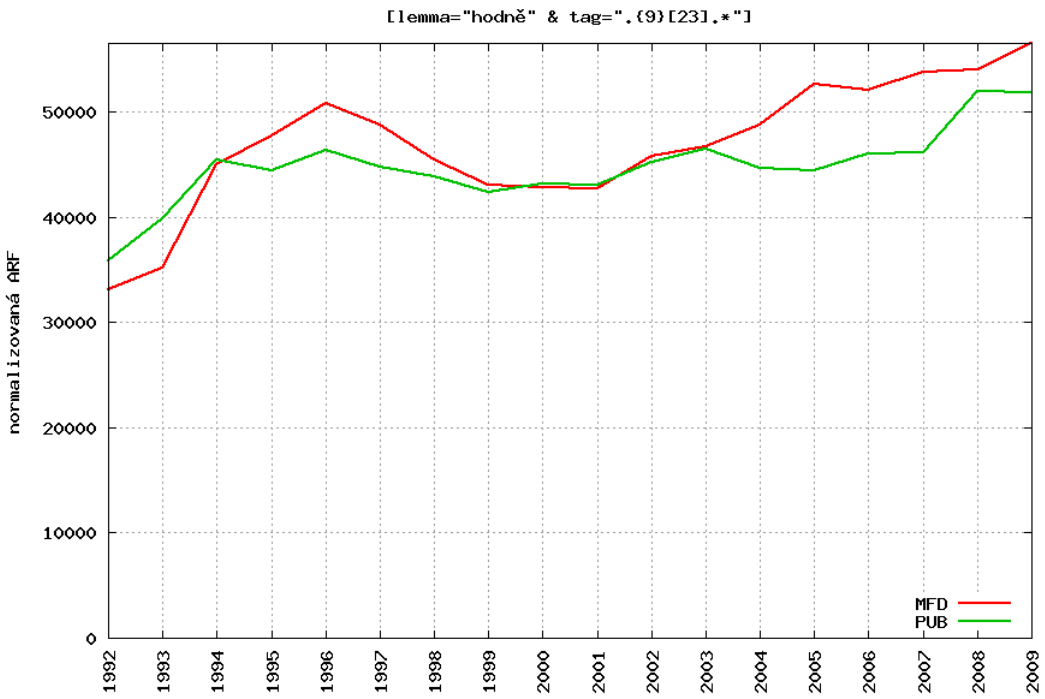
[P]*pramenící,* [M]*představitel* (obr. 6.4.8), [P]*příslušet*, [P]*rozporný*, [M]*tento souvislost* (součást spojení *v této souvislosti*), [M]*týž*, [P]*úloha*, [M]*ustavení*, [P]*vážný nebezpečí*, [P]*vnitřní rozpor*, [P]*vyjádřit přesvědčení*, [P]*vyřešení otázka*, [P]*vyslovený*, [P]*záležitost*, [P]*zásadně odlišný*.

Do této podskupiny řadíme i kombinace [P]*evropský integrační* (většinou jde o součást spojení *evropský integrační proces*), [P]*orgán státní* (součást spojení *orgány státní správy*), [M]*spolkový republika* (součást plného názvu Spolkové republiky Německo, který se dnes příliš neužívá), [P]*zákonodárný orgán*. Řadíme do ní také výrazy [P]*hromadný sdělovací*, [M]*sdělovací*, [M]*sdělovací prostředek*, [M]*tisk* (obr. 6.4.11), které jsou v současné publicistice nahrazovány pojmem *média*. Protože v použité verzi lemmatizace existuje pouze singulárové lemma *médium*, frekvence jeho singulárových tvarů je velice malá a tvar *média* je homonymní mezi sg. a pl., je průběhový graf na obr. 6.4.12 výsledkem dotazu na celé lemma.

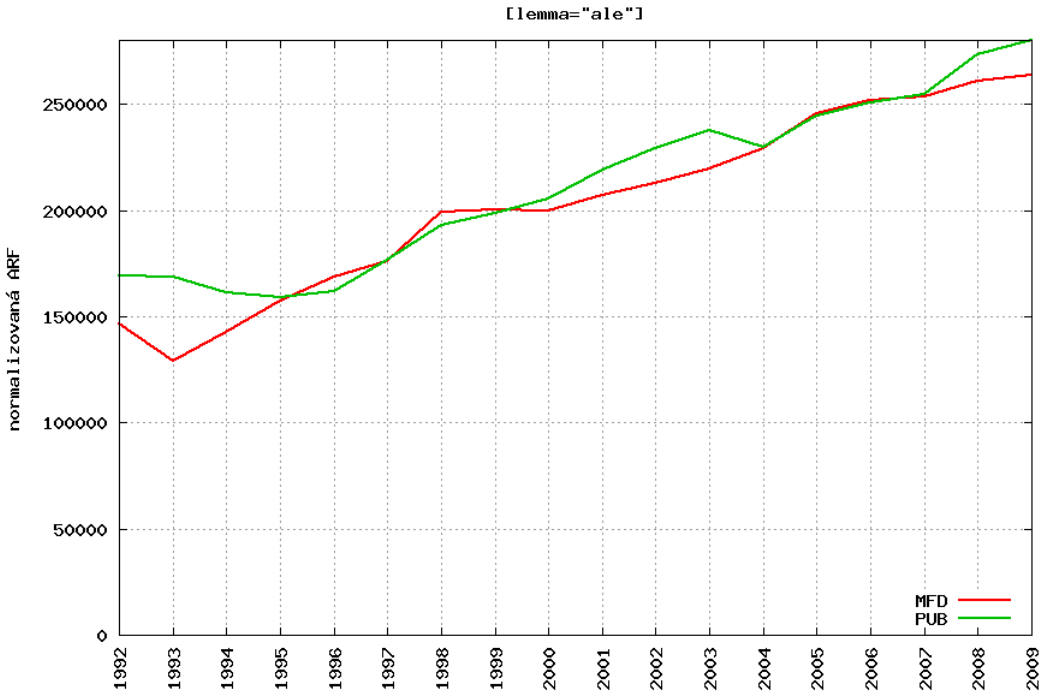
Celkově lze říci, že výsledky založené na *tau* jsou jazykově zajímavější než výsledky založené na iterativních *cbf*, *chi*, *ll* popisovaných v podkapitole 6.3. Poukazují nejenom na postupný nárůst neformálnosti jazyka publicistiky, ale také na množství šablonovitých slovních spojení využívajících přenesených významů sloves v kategoriálním užití. Právě tyto jazykové jevy jsou zvýrazněny metodou *taumed* a podrobněji popsány v následující podkapitole.



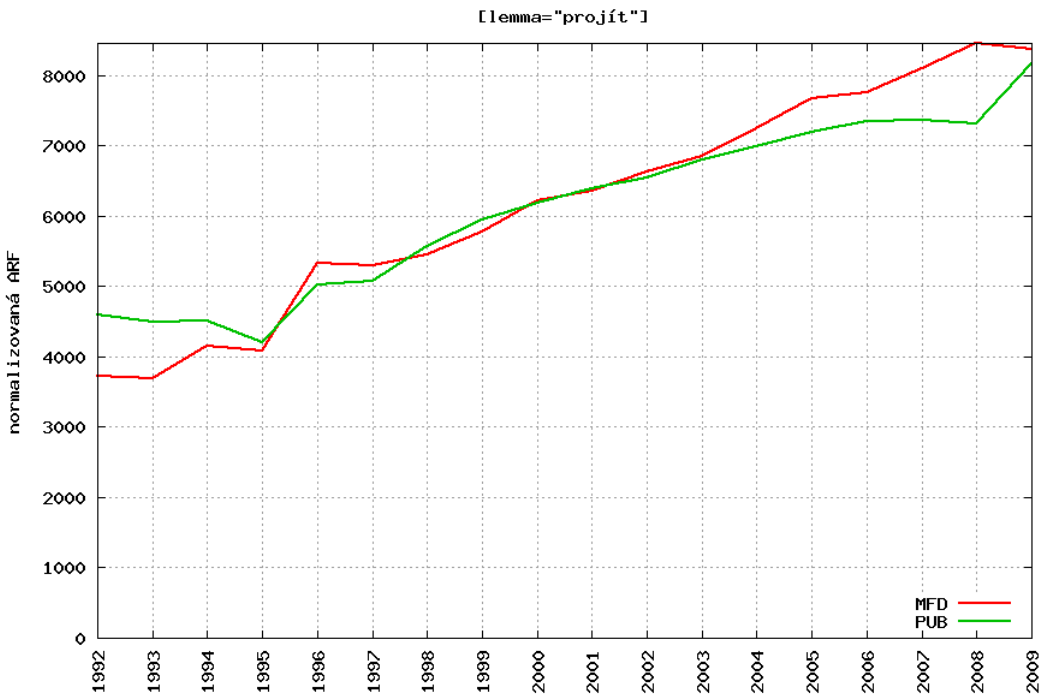
Obrázek 6.4.1: Průběh normalizované ARF pozitivu lemmatu *hodně*.



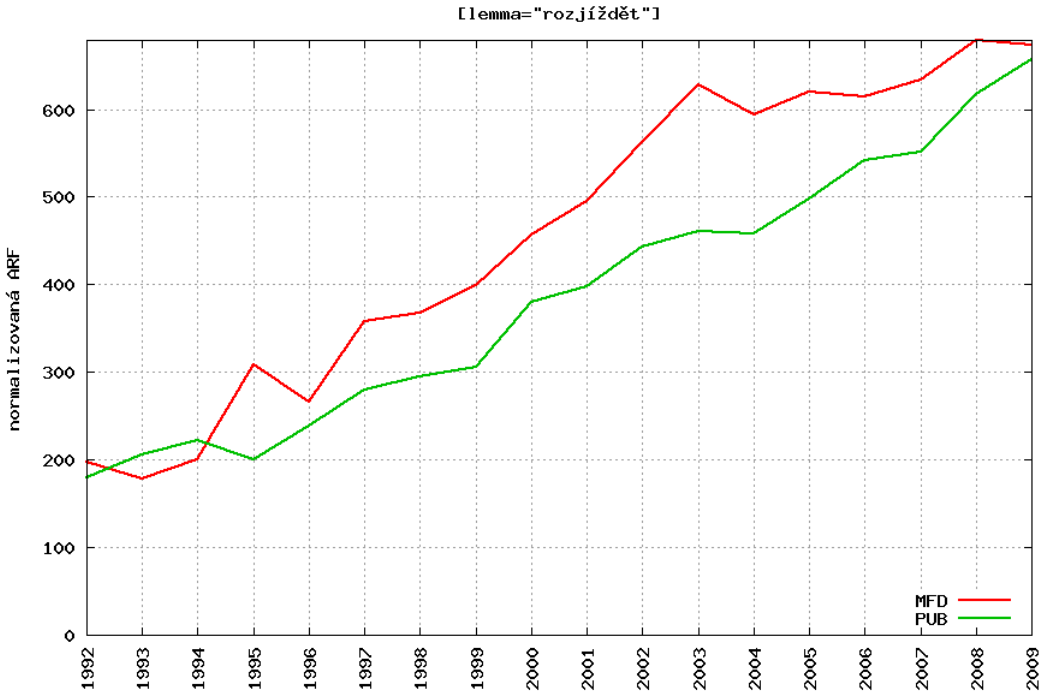
Obrázek 6.4.2: Průběh normalizované ARF komparativu a superlativu lemmatu *hodně*.



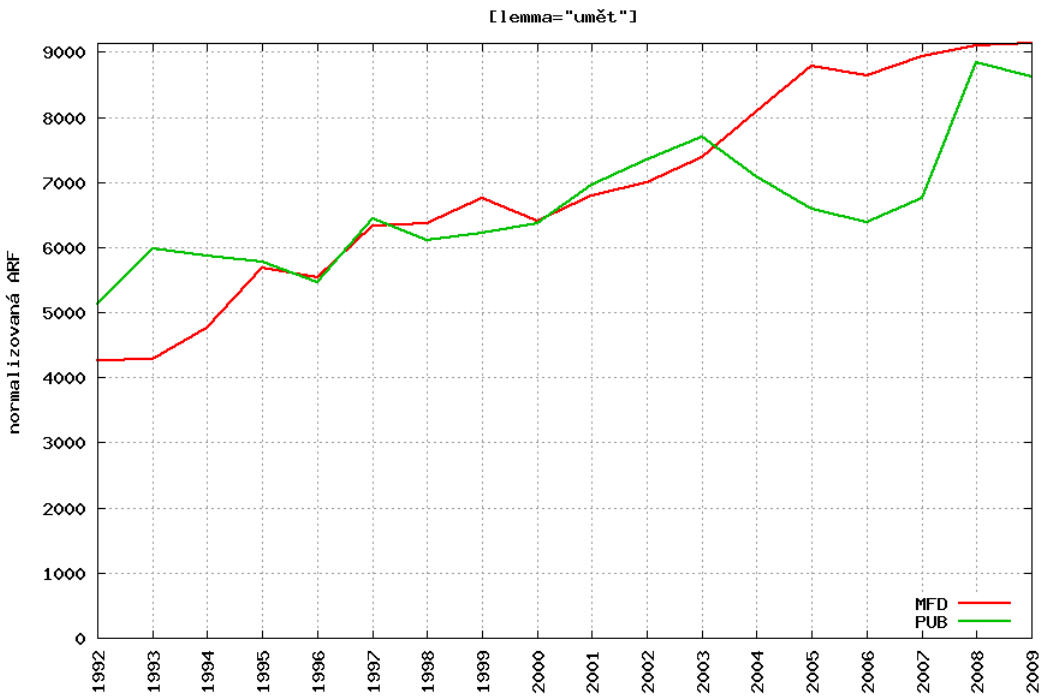
Obrázek 6.4.3: Průběh normalizované ARF lemmatu *ale*.



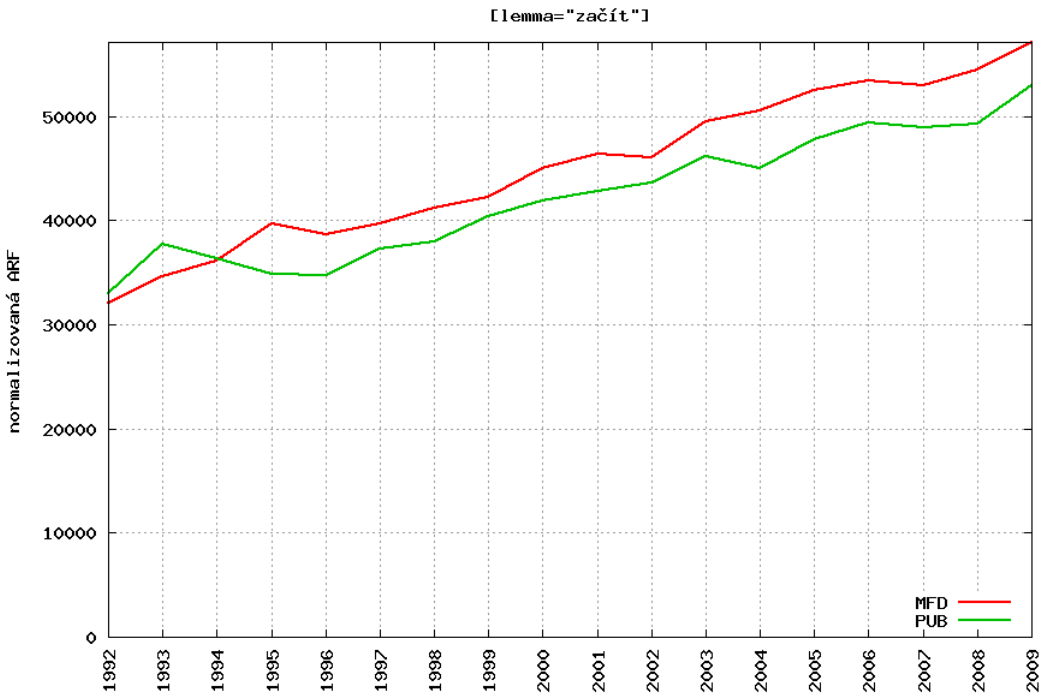
Obrázek 6.4.4: Průběh normalizované ARF lemmatu *projít*.



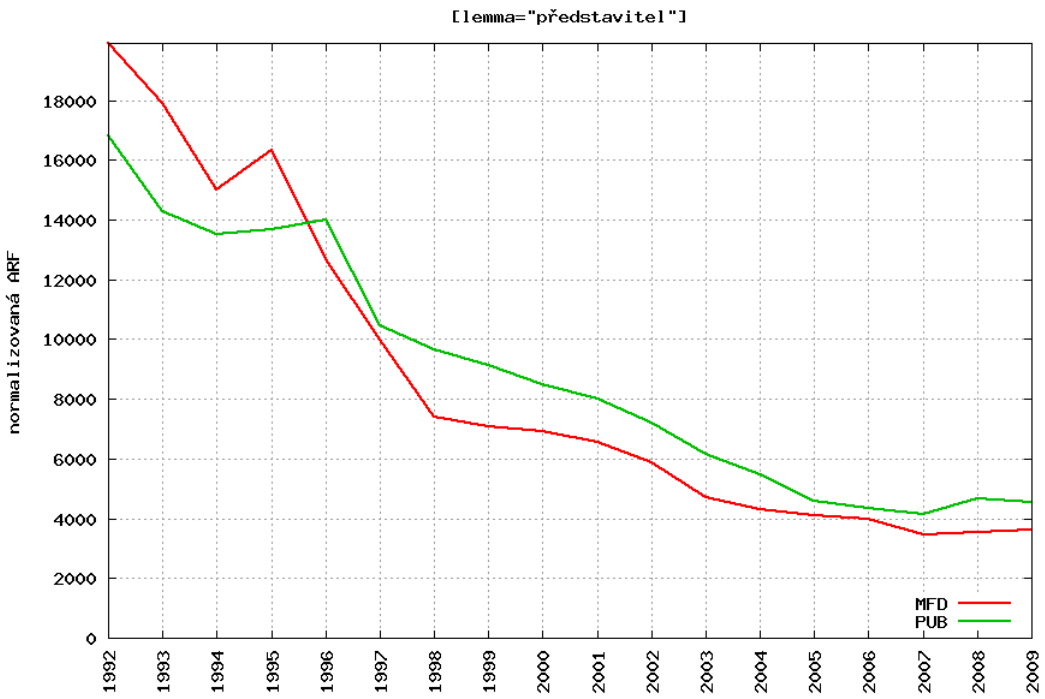
Obrázek 6.4.5: Průběh normalizované ARF lemmatu *rozjíždět*.



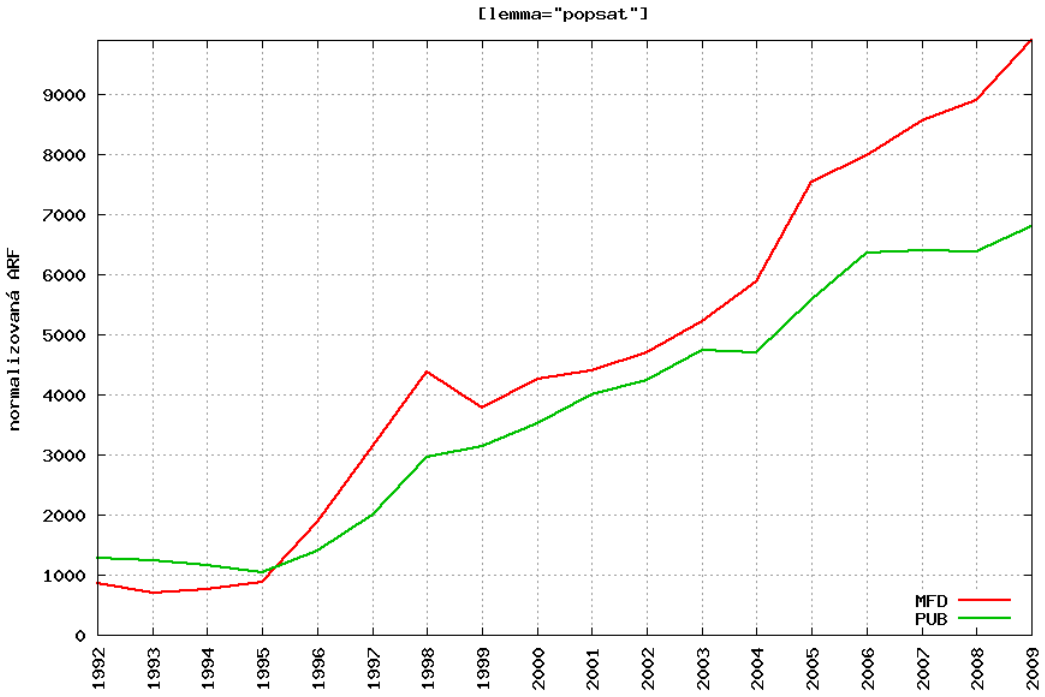
Obrázek 6.4.6: Průběh normalizované ARF lemmatu *umět*.



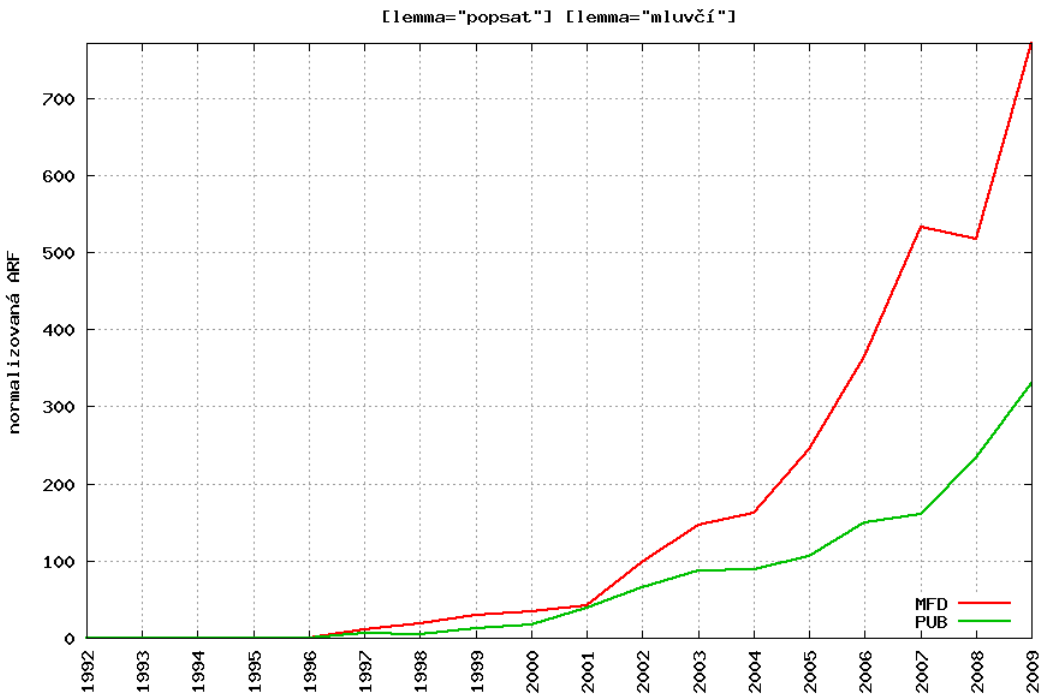
Obrázek 6.4.7: Průběh normalizované ARF lemmatu *začít*.



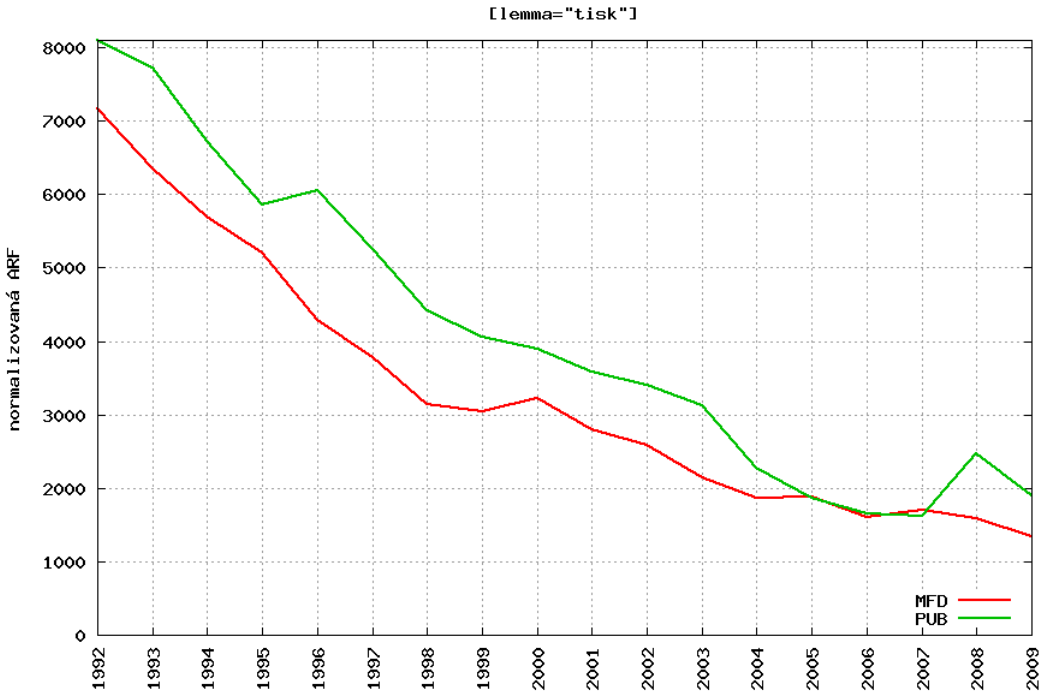
Obrázek 6.4.8: Průběh normalizované ARF lemmatu *představitel*.



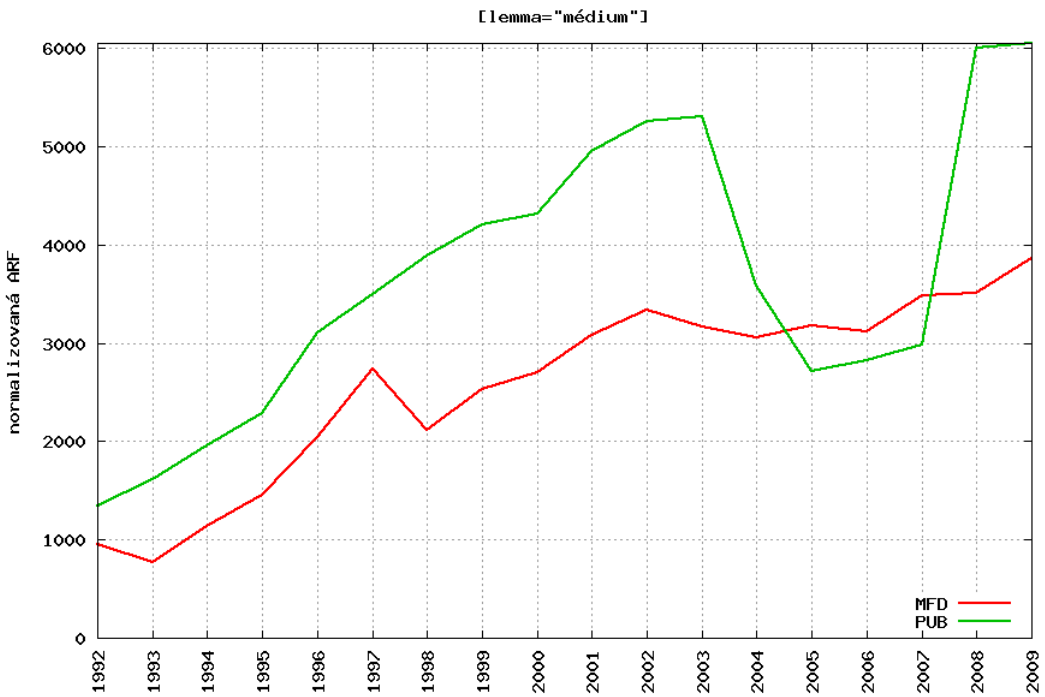
Obrázek 6.4.9: Průběh normalizované ARF lemmatu *popsat*.



Obrázek 6.4.10: Průběh normalizované ARF kombinace *popsat mluvčí*.



Obrázek 6.4.11: Průběh normalizované ARF lemmatu *tisk*.



Obrázek 6.4.12: Průběh normalizované ARF lemmatu *médium*.

6.5 *taumed* na publicistických subkorpusech

Tato podkapitola se zabývá rozborem výsledků metody *taumed* aplikované na publicistické subkorpusy řady pub_RRRR a mf_RRRR. Tato metoda vznikla empiricky jako korekce *tau* zvýhodňující frekventované výrazy, a to vynásobením jeho hodnoty desátou odmocninou mediánu frekvenčních hodnot (viz oddíl 5.3.2). Výrazy ve výsledných tabulkách by tedy měly vykazovat pravidelnou vývojovou tendenci a zároveň být dostatečně frekventované. Je tedy pochopitelné, že se řada výrazů z předchozí části opakuje, celkem jde o 17 lemmat a 18 kombinací. Jejich průběhové grafy jsou nejenom pravidelnější, ale zpravidla také výraznější (s vyšším nárůstem nebo poklesem) než u výrazů „doplněných“ pomocí *taumed* na základě jejich vysoké frekvence. Nárůsty a poklesy některých výrazů jsou tedy pozvolné, přesto však pravidelné (např. předložka *na* na obr. 6.5.15).

Na lexikální úrovni je jasně patrná převaha lemmat, která vykazují frekvenční nárůst, oproti lemmatům vykazujícím pokles. Celkový počet různých lemmat v obou tabulkách je 76, z toho 61 zaznamenalo nárůst a 15 pokles; mezi lemmaty vykazujícími pokles je navíc jediné, které najdeme v obou tabulkách ($^{\mathrm{MP}}$*však*). Srovnání s *Frequency Dictionary of Czech* (FDC, Čermák, Křen et al., 2011) ukazuje, že jde o velmi frekventovaná lemmata: z celkem 76 lemmat jich mezi nejfrekventovanějšími 100 lemmaty v FDC najdeme 43. Zjištěná převaha nárůstu frekventovaných lemmat je zajímavá i teoreticky, pokusili jsme se proto automaticky vygenerovat průběhové grafy pro všech 100 nejfrekventovanějších lemmat z FDC a vyhodnotit zjištěné vývojové tendence v publicistice. Výsledky potvrdily, že nárůst (byť nepravidelný) u nich převažuje nad poklesem v poměru 70:15, u zbylých 15 lemmat není celková tendence zřejmá. Zjištění příčin tohoto jevu by však vyžadovalo podrobnější analýzu, spokojíme se proto se závěrem, že převaha nárůstu nad poklesem v publicistice je typickým rysem frekventovaných lemmat a že není způsobena použitou metodou.

Zmiňovaná převaha frekvenčního nárůstu nad poklesem se však netýká lexikálních kombinací, kde pokles dokonce mírně převažuje (v poměru 48:52). Vzhledem k těmto faktům byly výrazy z výsledných tabulek pro další diskusi rozděleny do následujících tří skupin:

1. všechna lemmata; vzhledem k povaze použité metody jde o většinou běžná, frekventovaná lemmata z jádra slovní zásoby;
2. lexikální kombinace, které tvoří jmennou frázi (případně jsou její součástí);
3. ostatní lexikální kombinace.

Všechna lemmata 1. skupiny budeme nyní probírat po slovních druzích. Jedná se pouze o hrubé rozdělení pro větší přehlednost diskuse, při níž jsme museli respektovat fakt, že

rank	lemma		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	se	+	1073930	1100728	1096325	1112884	1086420	1164205	1154033	1153956	1180115	1197470	1217136	1240719	1257070	1324197	1322561	1338481	1334692	1346798
2	na	+	856644	887811	898644	944069	922541	951541	945225	962083	986145	992407	1001872	1015980	1035170	1084786	1087905	1089616	1044123	1058194
3	i	+	284637	278649	284787	291632	295851	299804	291750	293841	296579	300134	307744	311658	321636	328340	326112	325533	336477	350595
4	ale	+	169181	168638	161133	159278	161892	176722	192729	198378	205644	218952	229355	237961	229929	244615	250740	254398	273614	280144
5	do	+	293136	302335	302844	301805	310242	318267	332402	336902	346364	340965	346377	347511	370380	384447	385477	380536	352964	352723
6	o	–	387426	380029	383164	414762	381588	364783	352890	349266	348178	342111	338300	339289	321044	316826	312441	311495	331710	334828
7	hodně	+	42641	47256	52994	51241	53200	53869	54604	53822	55679	56698	59080	61482	61654	64661	67805	67967	71270	72110
8	tak	+	103640	106814	104417	107460	107862	115807	114138	111605	114353	119159	124718	126247	117528	123876	125189	133851	143578	144042
9	pak	+	42538	44273	45384	44969	46359	49229	48304	47913	49999	51458	53764	56442	53613	57694	57713	60742	66226	66525
10	další	+	78214	78972	81000	85716	88229	87326	90425	91834	92688	93443	93538	94036	100761	107688	106090	105827	96393	98667
11	oni	+	106480	111208	116481	107094	108127	110495	111511	114816	117254	120731	123600	127086	123893	127315	126597	124056	133336	137077
12	už	+	94187	95188	96168	80165	86522	91492	94563	99157	106479	113621	123537	129827	121297	127256	134537	134952	147876	153135
13	mít	+	301706	326500	335264	341466	328527	329284	331070	329078	335880	344128	347663	357121	340638	348023	351717	346391	369880	375924
14	od	+	121136	127424	129543	129845	130465	131814	134151	137660	139446	137220	139748	144360	152080	152214	151990	149733	143122	145056
15	také	+	61547	61252	69103	74051	79414	78650	82588	82345	83679	83735	84887	86984	93123	102521	97438	99986	90273	91714
16	k	–	261130	245413	240927	265735	239283	241407	232774	227353	225306	222077	215703	215323	211524	213193	207137	208939	215442	216646
17	tento	–	185334	177206	189406	215561	200253	183564	169707	156365	147325	143399	135257	128413	128512	135111	126685	136911	121920	120208
18	hlavně	+	9664	10557	11105	11176	11189	12739	13661	13618	14380	15047	15415	17065	17817	19602	20566	20851	20773	21719
19	moci	+	145186	149387	158449	154829	155740	159925	158587	156775	158824	163133	167084	168940	162424	162490	163067	166152	180736	183643
20	začít	+	32987	37705	36385	34821	34775	37248	37988	40393	41928	42793	43547	46136	44999	47748	49336	48877	49284	52987
21	dostat	+	35069	39145	40356	32030	37199	39296	43334	43710	45179	46776	49961	49190	48862	51531	54429	51903	52419	56039
22	kde	+	42679	44174	45221	40752	44194	47131	48663	48626	49326	50290	51394	52960	52863	54408	54043	55594	54594	54000
23	mezi	–	64782	62657	62486	63068	63286	60035	56835	56110	55857	55794	54356	54389	52812	52281	51948	51537	53810	52645
24	místo	+	46137	48516	49000	43310	50341	49855	54775	55664	57613	56514	61175	61492	68779	78757	80505	81877	63090	62282
25	přijít	+	26006	28478	28002	22593	25197	29727	31995	33588	34479	35610	38753	40600	42076	45298	45714	46313	43190	46072

Tabulka 6.5.1: Úroveň lexikální, *pub_RRRR*, **taumed**, 1. část.

rank	lemma		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	přes	+	24716	27449	28002	28104	29011	28710	28697	30045	30115	29805	30729	31841	30239	30361	32318	32732	32670	35346
27	pomoci	+	9789	10533	11207	10746	10741	12217	13139	13875	13856	14238	15570	15255	15665	16513	16826	16895	17359	17477
28	navíc	+	15529	19066	20653	16558	21893	22789	24812	24778	24322	24745	25460	26175	25635	26646	27482	27057	27790	28152
29	muset	+	71672	75191	79197	74387	77862	81578	81204	79601	83837	85532	88149	89051	84249	86681	87934	86186	89881	92613
30	po	+	158905	157353	163552	150271	162350	165410	168888	165072	167565	166784	171518	168927	178493	192507	191983	190752	173580	179768
31	celý	+	49357	50239	50431	51917	52372	55041	54586	54828	54922	55963	59920	60642	56957	60054	58881	61259	59335	60370
32	nechat	+	10109	11453	10904	10445	11238	13754	14385	14098	14951	15400	16412	17102	17725	19083	19230	19857	20659	19617
33	díky	+	8374	10372	11305	11023	11738	12070	11980	11935	12951	12527	12949	13512	13758	14916	15872	16671	18168	18198
34	ještě	+	63736	63102	62579	59750	62226	64816	66639	66296	68454	68539	71648	72066	69307	72372	72511	71895	72332	74833
35	tři	+	44104	48984	52392	49438	54404	52599	53767	56256	57773	56620	58096	58348	61200	66175	67524	65333	57843	60978
36	německý	–	21063	20645	20029	19946	19783	16584	15272	14799	14463	13990	12811	12215	9978	8489	7576	6864	9432	9435
37	však	–	99223	108167	129198	109149	135376	124027	111677	103017	97171	89450	84909	80971	71780	68094	63826	58139	60422	61523
38	najít	+	15518	15754	15602	13877	14712	17779	17933	19253	19502	20544	21515	22950	22602	23802	24103	24986	24553	24936
39	jeho	–	102361	100090	102278	95593	102161	101383	97140	92319	92369	93875	90329	89353	80189	77456	74806	73220	90488	89865
40	mluvčí	+	7957	9186	9809	9625	14618	12055	14312	15804	17671	18551	19356	20823	17914	18784	21765	21460	21494	22365
41	chtít	+	60717	67010	67458	58545	66838	67857	72867	72554	75201	78858	83451	85598	78050	76340	78019	81903	90665	92271
42	předseda	–	30180	25750	23580	20628	25780	19424	20235	15940	14829	13834	14438	11931	11608	11215	11533	9891	10886	11208
43	před	+	74788	76451	81485	68417	80930	80330	85497	83362	85149	85171	89437	86031	87104	88415	91543	87024	88554	88561
44	dokázat	+	13968	15205	15173	14326	15953	18088	18189	17758	18142	18458	18519	18559	19395	21225	21400	21869	21468	20605
45	čekat	+	15637	16708	16934	14168	16458	19469	21347	22144	23351	24599	25288	26378	28819	31655	33356	33939	28157	28368
46	vláda	–	50300	40862	38161	40386	37218	35419	32300	26178	20613	18382	20207	17881	10800	5566	6135	5540	15994	16134
47	plánovat	+	1897	2312	2540	3358	3644	3368	4038	4355	4657	4904	5040	5314	5809	6244	6236	6631	6903	6915
48	nakonec	+	13578	13817	13943	10667	12869	15150	16324	15946	16800	17303	17900	18016	17922	20038	20298	20737	19635	21213
49	prostředek	–	14813	14193	14423	17768	14018	12456	11563	10223	9411	9271	8877	7984	7870	7685	6886	6820	6183	6965
50	jednání	–	26120	22269	19856	19951	19602	16043	15906	15107	14636	14001	13275	11962	11135	10837	10957	11303	11202	11132

Tabulka 6.5.2: Úroveň lexikální, *pub_RRRR*, **taumed**, 2. část.

rank	lemma		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	se	+	941137	903909	1006130	1073388	1130810	1178170	1144855	1117537	1139727	1152343	1175994	1208757	1251770	1328958	1319009	1335606	1340933	1342387
2	na	+	896710	879832	901835	911812	930524	933589	914448	920465	950829	972427	1002630	1025165	1059376	1094527	1098151	1095754	1085657	1107623
3	ten	+	422592	401047	423880	462500	475962	504967	481416	463331	469732	499325	512797	526612	552580	608522	623875	646804	680729	703559
4	ale	+	146391	128785	142360	157165	168729	176072	199393	200345	199501	207181	213048	219320	229121	245405	251706	253641	260936	263500
5	už	+	90426	83934	103565	106596	105848	109639	118474	115100	116248	121684	132092	140755	145164	158148	162784	166995	170996	175298
6	chtít	+	62391	63530	70665	75766	76565	77136	80183	77777	78815	81949	86452	89728	94150	98718	97777	101445	104105	104838
7	začít	+	32094	34575	36079	39695	38620	39716	41214	42254	44990	46401	46032	49455	50526	52494	53450	52937	54491	57100
8	být	+	1544011	1475588	1555028	1653096	1626744	1594409	1545188	1491334	1540788	1572654	1602478	1629948	1713737	1837222	1856979	1879012	1901041	1924157
9	rok	+	171835	184117	195962	227429	196587	196614	202563	207369	210085	212690	208729	213539	216441	237755	232283	237153	236478	239929
10	člověk	+	53081	55294	67207	70744	73279	89536	98268	97152	100573	102500	110415	122770	136852	142646	135215	132680	130072	140710
11	čekat	+	17308	16360	17685	18996	19062	22908	25338	25434	26499	28126	29344	30726	32184	32855	33129	33244	34196	33525
12	i	+	272953	254826	269462	286688	305054	314458	293953	288443	289068	292938	301154	307955	335752	361862	368891	376569	374552	375252
13	vědět	+	24891	21999	25040	27293	29613	32440	33802	32873	33848	34487	33849	35523	36017	37651	37189	37702	38528	39706
14	tam	+	18862	16880	16804	18254	20257	23075	27081	27476	28818	30872	32644	35092	37215	40880	42084	43668	43941	46231
15	co	+	74880	62770	69161	71399	73336	81516	76802	77317	75463	81382	81395	82252	86554	90804	92749	96971	99948	100197
16	do	+	302525	299807	303146	297343	315643	321965	351345	351726	354961	351666	360594	364005	370996	367416	366739	366107	364881	364936
17	oni	+	95521	91762	109385	118693	119838	123240	119726	118171	118188	119929	122521	126250	133950	143076	140924	146201	149947	152894
18	od	+	126043	137800	134994	141925	134361	134734	140739	141349	142423	141341	149460	154231	158286	154872	157719	161523	162933	162126
19	hodně	+	39850	41104	51874	54586	59054	58279	58297	57354	57989	58992	61710	63607	67104	73829	73603	76097	78216	80304
20	dostat	+	37172	39379	45893	43582	44880	45492	47223	45882	46151	48034	52222	51169	52607	55886	56018	54877	56853	60229
21	republika	–	46344	53328	44765	36333	28037	24956	23661	23462	22726	23372	23069	22932	21419	20390	19858	19693	18996	18207
22	místo	+	44565	48320	50961	51399	57135	55982	61747	60793	61750	60508	68239	68968	70953	72618	74705	72039	71943	73612
23	kde	+	37949	38953	43799	45722	48255	50426	51762	49769	49566	51506	53429	56278	58612	59863	59577	60506	59842	60449
24	tak	+	87800	84008	92441	101138	110221	116327	108218	102078	104285	108067	111376	113190	115096	124234	126248	128899	133411	139597
25	najít	+	14751	13578	14335	15546	16812	19001	19730	20358	20269	20746	21813	23307	23830	25605	25898	26337	26177	26637

Tabulka 6.5.3: Úroveň lexikální, *mf_RRRR*, **taumed**, 1. část.

rank	lemma		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	také	+	60077	60544	72125	77294	84243	83333	82086	79500	81543	82743	86212	89590	92991	96210	97787	96316	97022	97292
27	hlavně	+	9172	9831	11532	12489	12725	14010	15839	15174	15642	16517	16697	18689	20157	21682	22201	23805	24927	24655
28	přijít	+	26877	26395	29636	29258	31653	34252	36335	36175	36014	36966	40874	43440	45852	47891	46568	47790	47643	51142
29	třeba	+	22179	20756	19639	23581	23937	24304	23517	24795	24747	25500	26860	29314	31317	34454	37176	42000	44041	45256
30	on	+	141227	148206	177225	196730	191299	189678	180898	172609	176911	188248	191652	202531	210071	232044	235739	238648	243545	246267
31	říkat	+	23544	23835	28412	29215	28596	32942	39282	41270	39693	43023	46342	51146	56622	62009	61260	59161	57388	58200
32	snažit	+	14147	13411	15226	16900	16863	17461	18247	18101	18871	18798	19161	18963	19325	20263	20181	20549	20646	20794
33	pomoci	+	10070	10387	11425	12489	12401	13629	14330	14492	14422	14830	16637	16477	17054	18795	18401	18795	19132	19898
34	člen	–	25081	24002	23902	24324	21846	21309	22170	19521	19276	17834	17794	15143	14785	14312	13876	13117	13455	13358
35	země	–	49661	45463	46408	41617	40628	43294	31634	28761	28979	28553	26235	26649	25848	26035	26375	25005	24262	24259
36	jestli	+	5061	5249	5938	6594	7608	8546	8371	8565	8861	9342	9404	9948	10905	12315	13103	13397	14693	15645
37	pak	+	39487	38433	42296	47294	49724	50887	46163	44913	47453	48749	50798	53931	55632	61014	62789	65476	67403	67963
38	když	+	59748	56444	66294	72141	73610	80252	80818	78524	77090	77356	76227	80751	82210	87771	87754	89331	92559	93483
39	však	–	103674	125131	155546	169567	169116	161162	116276	98877	94867	90911	88672	84897	77978	79484	71108	68092	69453	70301
40	teď	+	13611	11000	10136	8952	11123	13970	15204	16762	18302	20410	21874	25089	26125	28157	31093	32569	34808	38633
41	chystat	+	4284	4563	5691	5328	6464	6647	7106	7380	8186	8525	8741	10308	11034	11247	11735	12068	12459	12561
42	dělat	+	14544	12892	14141	13625	16532	19711	21447	21336	21053	21957	22840	23146	24697	25443	25991	25626	26379	27296
43	zase	+	8568	8588	10190	12140	12121	13462	13556	13363	13970	13627	14596	15653	15658	16298	17598	18123	19010	18912
44	dát	+	39193	36579	38913	38778	42243	43323	44292	42129	42533	42884	44682	44566	46507	49502	49346	49740	50375	52137
45	udělat	+	11383	10944	12423	12271	14721	16710	17891	17130	16897	17280	18301	18199	19624	20225	20460	20925	22451	22774
46	pár	+	8257	6641	6829	8079	8339	9890	10075	10478	11296	11275	11751	12933	13939	14988	15269	16116	16344	16677
47	kvůli	+	8585	11129	17449	22795	25596	26589	23818	23669	24286	26413	29140	31112	30963	33040	34728	34124	34995	41512
48	díky	+	7531	8199	10738	10917	11600	11563	11752	11443	12892	12383	13246	13598	14274	16433	17262	17955	18667	20070
49	zahraniční	–	25478	20682	17814	20262	15007	13133	9036	8524	8367	8170	7614	7455	7553	7211	6800	6525	6362	6090
50	hned	+	7928	7679	7978	8821	9935	11708	13261	13621	13252	14196	15037	15579	15182	16036	16597	17585	17974	17580

Tabulka 6.5.4: Úroveň lexikální, *mf_RRRR*, **taumed**, 2. část.

rank	kombinace		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	tisíc koruna	+	2954	4845	8542	5101	7822	8024	9658	10575	11574	11511	13198	14511	17519	22020	22794	21048	11581	12652
2	souvislost s	–	6759	6550	6179	6953	5576	5241	4791	4546	4331	4365	4019	3665	3343	2872	2763	2770	3565	3407
3	uvést mluvčí	+	125	168	312	346	1018	926	1201	1596	1985	2290	2762	3007	2509	2470	3140	3463	3060	3255
4	webový stránka	+	0	0	0	0	24	43	114	290	654	809	886	1105	1142	1290	1621	2198	2364	2280
5	akciový společnost	–	3344	3463	4022	3985	3463	2195	2089	1799	1684	1558	1165	986	949	1095	814	730	670	494
6	doufat že	+	3735	3660	3514	3121	3867	4166	4714	4611	4669	4932	5120	5038	5178	5598	5646	5738	5158	5192
7	považovat za	–	9892	9944	11086	11042	11207	10093	9523	8597	8042	8230	7918	7398	5612	4390	4361	4069	6233	6053
8	o víkend	+	2054	2099	2204	1699	2689	2947	3832	4275	4817	4990	5101	5569	6946	8489	8444	8539	5260	4850
9	tento souvislost	–	3035	2497	2670	3012	2579	1968	1753	1427	1385	1181	1041	916	762	631	486	523	607	570
10	od začátek	+	2271	2630	2852	2632	2859	2733	3127	3303	3213	3181	3150	3331	3603	3893	4382	4302	4122	3875
11	tiskový konference	–	5680	4608	3066	3728	2673	2124	2377	2007	1919	1615	1557	1502	1107	952	906	866	1340	1596
12	spolu s	–	8483	8701	8733	8188	9196	8897	8870	8401	8393	7980	7638	7757	7453	7155	6709	6817	7131	6750
13	patřit mezi	+	1388	1810	1943	2025	2236	2328	2322	2301	2298	2399	2434	2783	2662	2647	2540	2674	2870	2748
14	sdělovací prostředek	–	2867	2867	2484	1970	1911	1535	1287	958	738	677	615	483	319	213	165	203	278	241
15	žádný případ	–	2374	2151	2260	2188	2081	2161	2252	1959	1892	1983	1866	1634	1548	1707	1671	1429	1378	1304
16	vzhledem k	–	8407	8157	8271	9408	9042	8316	7929	7095	7072	6650	6444	5830	6405	6995	6620	6537	5639	5876
17	přesvědčení že	–	1767	1526	1347	1244	1220	1194	870	738	715	688	614	600	363	265	231	222	468	393
18	rozpor s	–	2125	2203	2223	1926	1832	1552	1611	1503	1436	1501	1290	1262	801	686	615	611	1049	899
19	státní správa	–	2174	2411	1836	2025	1550	1528	1722	1519	1509	1678	1469	1105	798	605	605	521	556	621
20	policejní mluvčí	+	81	104	154	94	427	564	1032	1717	2355	2539	2691	3119	2748	3642	4638	4116	2251	2900
21	národní majetek	–	2190	2642	2945	2178	2456	1209	1314	1014	982	748	655	457	220	113	72	25	51	76
22	vztah mezi	–	2829	2191	2395	2761	2223	1983	1522	1455	1350	1265	1149	993	744	563	509	513	834	823
23	celý záležitost	–	1257	1451	1226	869	997	926	976	945	778	818	612	558	552	598	494	464	417	355
24	pár den	+	889	942	871	538	681	1004	1075	1185	1217	1260	1537	1557	1357	1368	1734	1678	1795	2014
25	soulad s	–	2564	2133	2125	3101	2325	1977	1770	1652	1749	1817	1542	1465	1232	1168	1026	1093	1226	1140

Tabulka 6.5.5: Úroveň lexikálních kombinací, *pub_RRRR*, taumed, 1. část.

rank	kombinace		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	dospět k	–	1897	1607	1589	1625	1437	1257	1159	1065	1055	1035	949	909	656	527	484	513	961	545
27	domnívat že	–	4342	4833	6305	4998	5613	4432	4643	3849	3187	3308	2766	2586	1847	1521	1356	1248	1960	1697
28	centrum město	+	786	815	979	602	1097	1133	1388	1431	1579	1674	1985	2147	2313	2534	2603	2605	1821	1836
29	rodinný dům	+	92	208	331	370	293	455	461	597	691	784	1069	1268	1480	1949	2020	2125	1201	1659
30	zlínský kraj	+	0	0	0	0	0	6	18	32	309	461	641	679	756	848	884	1462	999	937
31	základní škola	+	1046	1486	1407	1002	1435	1662	2305	2788	3114	3043	3345	3985	6998	9650	9150	8999	3527	3913
32	směřovat k	–	1122	919	792	894	830	780	741	741	717	715	659	668	545	523	522	478	556	532
33	podstatný část	–	650	711	657	588	618	550	573	508	498	497	485	445	373	318	333	264	329	304
34	tři bod	+	228	295	587	316	552	594	869	962	1056	1102	1070	1057	2056	3158	3386	3634	1631	1786
35	ministr vnitro	–	3328	3544	2838	1827	2589	2260	2105	1648	1813	1775	1618	1311	973	598	637	608	1290	1064
36	hned několik	+	222	295	322	341	372	509	617	679	616	646	708	785	929	1154	1204	1336	1151	1076
37	k vidění	+	699	647	680	548	1157	1374	1512	1740	2055	1911	2007	2175	3306	3957	4031	3921	2642	2875
38	otázka zda	–	2623	2480	2684	2015	2212	2273	1948	1720	1566	1507	1559	1404	1035	836	758	800	1037	1254
39	předseda vláda	–	2358	1914	1589	1561	1654	1370	1097	973	875	669	721	691	456	421	337	240	620	519
40	ředitelství silnice	+	0	0	0	10	26	76	220	256	336	454	400	480	442	576	699	837	620	709
41	stránka www	+	0	0	0	0	0	19	44	142	322	515	644	680	638	679	787	1098	1113	1026
42	ministr zahraničí	–	4656	5018	4930	4114	3689	2857	2158	2123	1628	1803	1645	1674	780	278	289	302	1454	1266
43	záchranný služba	+	455	480	555	227	508	592	826	869	1040	1078	1327	1420	1842	2242	2199	2327	1682	1507
44	označit za	–	2938	2798	2819	2751	3039	2365	2280	2035	1949	1956	1957	1836	1249	939	844	873	1871	1760
45	spor mezi	–	1003	948	960	785	861	719	712	684	581	676	573	540	309	288	269	264	506	431
46	střední škola	+	900	1041	1212	1022	1081	1147	1356	1607	1630	1723	1639	1845	2169	2828	2891	2958	1808	1659
47	přispět k	–	2179	2162	2358	2662	2382	2256	2136	1952	2045	1881	1716	1683	1681	1732	1614	1665	1682	1583
48	značný část	–	856	931	829	844	777	800	610	555	597	503	493	441	350	298	276	280	253	355
49	metr čtvereční	+	434	567	746	815	1128	1014	999	1007	1108	1191	1234	1256	1289	1340	1335	1592	1163	1431
50	státní rozpočet	–	2640	3145	3542	4474	3895	3691	2617	2453	2370	2118	1983	1652	1196	947	1164	960	923	1444

Tabulka 6.5.6: Úroveň lexikálních kombinací, *pub_RRRR*, **taumed**, 2. část.

rank	kombinace		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1	myslet že	+	7117	6882	7441	8734	8930	10571	11149	10757	11066	11151	11271	10521	11347	12202	11653	11546	12470	12630
2	spolu s	–	8982	8013	8719	8777	9045	8523	8478	7625	7857	7616	7412	7257	7457	7254	7170	7046	6767	6419
3	uvést mluvčí	+	242	334	483	873	2301	2343	2139	2331	2909	3163	3750	3707	3568	4114	4480	4226	4601	4907
4	považovat za	–	9794	9108	10641	11005	10640	9596	8148	6829	6655	6706	6477	5663	5758	5721	5650	5233	5394	5387
5	internetový stránka	+	0	0	0	0	13	110	274	747	1367	2097	2447	2363	2534	2558	2738	3111	3063	2578
6	k vidění	+	881	575	655	830	1379	1714	2139	2519	2849	2580	2822	3315	3701	3515	3613	3871	3754	3655
7	úplně jiný	+	639	575	601	611	833	935	1114	1299	1110	1222	1305	1323	1366	1385	1385	1521	1665	1673
8	říci mluvčí	+	242	538	515	917	2148	2002	1719	1846	2240	2445	2687	3470	3394	3785	3722	3879	3692	4231
9	webový stránka	+	0	0	0	0	44	58	85	236	508	618	804	968	1147	1318	1697	2026	2404	2448
10	poté co	–	6944	9052	15247	13494	12719	11171	8345	6766	6299	6328	6554	6298	5338	5639	5501	5147	5056	5393
11	oba strana	–	5078	3747	3812	4629	3305	3012	2408	2272	2282	2257	2161	1756	1846	1908	1784	1744	1821	1756
12	pár den	+	950	686	752	655	909	1039	1140	1339	1336	1313	1521	1584	1804	1817	2226	2297	2444	2520
13	akciový společnost	–	5027	5249	4542	3101	2562	1933	1664	1308	1354	1381	1085	938	797	1076	730	721	833	665
14	minulý týden	–	6806	7828	8182	7904	7297	7068	5788	5826	5529	5210	5335	5300	4852	4928	4937	4843	4830	5051
15	zlínský kraj	+	0	0	0	0	0	6	56	87	633	604	697	771	835	1028	1028	1089	1188	1250
16	celý záležitost	–	1676	1948	1235	1572	1182	1004	702	700	559	664	513	463	434	392	342	332	262	329
17	ministr vnitro	–	4163	3988	3125	3013	2301	2077	1310	1072	1079	1033	990	835	966	913	986	824	885	757
18	začít stavět	+	207	241	505	349	451	531	603	603	689	728	868	1008	971	1213	1292	1226	1284	1342
19	liberecký kraj	+	0	0	0	0	0	6	13	13	300	848	888	1062	1127	1339	1330	1344	1350	1391
20	národní majetek	–	3178	3580	3790	2838	2549	1506	1135	765	631	539	421	312	167	170	94	68	56	45
21	bytový dům	+	17	74	86	87	127	202	239	296	369	487	729	854	927	1036	948	1051	1090	1053
22	sdělovací prostředek	–	2695	2448	2169	2271	1678	1264	805	603	436	419	370	257	210	212	174	165	128	121
23	domnívat že	–	4422	5324	8268	8166	6566	5384	4454	3275	2699	2821	2079	1868	1690	1692	1695	1447	1292	1304
24	stavební povolení	+	466	371	472	218	578	814	595	740	839	917	971	1091	1056	1153	1167	1272	1246	1366
25	tento souvislost	–	3627	3228	2931	2926	1831	1368	811	615	725	658	569	429	440	456	395	344	351	327

Tabulka 6.5.7: Úroveň lexikálních kombinací, *mf_RRRR*, **taumed**, 1. část.

rank	kombinace		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
26	otázka zda	–	3023	2708	2620	2358	2167	2366	1624	1384	1204	1087	1172	1004	1011	957	923	890	957	970
27	ředitelství silnice	+	0	0	0	175	19	138	353	406	436	678	610	771	713	880	1013	1059	1146	1160
28	vědět jestli	+	397	556	591	786	814	1033	1002	1068	1045	1112	1031	1094	1248	1351	1334	1322	1396	1539
29	po skončení	–	2125	2077	2019	1834	1989	1933	1600	1638	1602	1601	1519	1454	1552	1478	1395	1357	1406	1422
30	sloužit jako	+	743	761	752	655	1087	987	1052	1113	1123	1186	1170	1183	1267	1296	1387	1439	1510	1373
31	fond národní	–	1399	2282	2556	2052	1913	1062	1013	660	573	500	387	280	167	176	90	68	66	40
32	věřit že	+	4577	4749	5154	6026	6375	6653	7829	7226	7284	7175	6846	6758	6883	7310	7453	7438	7860	8433
33	vztah mezi	–	2902	2003	1922	2183	1646	1673	1071	1014	989	956	798	648	661	606	627	655	627	618
34	doprovodný program	+	69	19	21	0	108	173	213	363	506	531	539	737	863	749	917	1125	1168	1183
35	střední škola	+	1002	1150	1213	961	1220	1235	1640	2032	2027	2116	1968	2193	2070	2513	2159	2410	2213	2448
36	rodinný dům	+	121	278	301	480	273	352	534	563	736	859	1160	1558	1647	1934	1746	1785	1679	2020
37	občanský sdružení	+	225	241	644	568	661	571	1137	1765	1909	2271	2233	2452	2431	2201	2084	2510	2492	2573
38	olomoucký kraj	+	0	0	0	0	0	23	24	121	416	728	691	790	812	959	1015	1009	1006	1050
39	ministr zahraničí	–	4180	5398	5358	5109	2708	2337	877	834	759	784	730	743	702	589	560	527	585	571
40	spojený stát	–	3783	3432	6314	5852	7068	5395	3014	2631	2641	3416	2674	2489	2164	2089	2113	1982	1891	1718
41	vůbec poprvé	+	380	408	494	524	706	727	670	727	797	836	864	769	852	933	1018	1019	1036	999
42	vysvětlit mluvčí	+	0	0	32	0	83	110	202	247	298	297	396	437	445	529	554	581	609	762
43	milión marka	–	1503	1410	1396	1572	1322	992	470	419	367	359	54	36	17	12	10	10	6	9
44	zahraniční firma	–	1244	1002	1020	611	788	462	439	431	376	351	259	304	281	275	258	213	154	130
45	pár minuta	+	345	185	236	349	356	421	505	559	553	531	597	752	696	832	804	858	966	907
46	stránka www	+	0	0	0	0	0	17	13	135	269	523	777	671	743	949	1062	1495	1655	1812
47	pardubický kraj	+	0	0	0	0	6	6	11	22	392	1064	1123	1195	1241	1359	1393	1282	1314	1279
48	valný hromada	–	1468	3135	2792	3319	2237	1696	1605	1418	1206	1068	672	543	507	569	415	430	365	419
49	tisíc marka	–	587	742	698	480	458	433	322	368	311	270	71	45	30	18	17	16	8	11
50	ministerstvo vnitro	–	4336	6715	4445	4716	2377	2245	1696	1851	1996	1858	1478	1329	1226	1179	1276	1234	1262	1153

Tabulka 6.5.8: Úroveň lexikálních kombinací, *mf_RRRR*, **taumed**, 2. část.

řada slovnědruhových distinkcí není v korpusech vůbec označena, což se týká zejména častého užívání adverbií v částicové platnosti ([P]*ještě*, [P]*navíc*, [M]*zase* atd.); určování částic je pro desambiguaci zvláště nesnadné už pro absenci obecně přijímaného teoretického konceptu. V jiných případech jsou sice výskyty daného lemmatu v korpusech označeny dvěma nebo třemi slovními druhy, ty jsou však při vyhodnocení spojeny do lemmatu jediného, protože použitá metoda nebere v úvahu slovní druh a drží se pouze formy lemmatu. To se z výsledných tabulek týká následujících lemmat (v závorkách jsou uvedeny rozlišované slovní druhy): [M]*co* (zájmeno, adverbium nebo spojka), [MP]*místo* (substantivum, předložka nebo spojka), [M]*pár* (číslovka nebo substantivum) a [MP]*tak* (adverbium, spojka nebo částice). Na příkladu lemmatu [MP]*místo* lze ukázat, že nerozlišování substantiva od předložky a spojky je jiného druhu, než nerozlišování v rámci těchto nesklonných slovních druhů: zatímco v případě substantiva *místo* jde (ze synchronního pohledu) o nesporně odlišný lexém, ve druhém případě slovní druh pouze vyjadřuje různou funkci daného lexému ve větě.

Jsme si vědomi toho, že zvolené rozlišování lemmat pouze podle formy, implementované již v základu použitých metod, je zejména z teoretických důvodů nevhodné. Domníváme se však, že je prakticky ospravedlnitelné řadou nepoměrně častějších případů druhého typu, kdy by – zejména kvůli obtížné desambiguaci – výsledky nebyly dostatečně spolehlivé. Při následné slovnědruhové kategorizaci proto budeme vycházet z toho, jaký slovní druh u daného lemmatu v korpusu převažuje. Tato kategorizace tedy vychází z výsledků morfologické analýzy a desambiguace, a proto také lemma [MP]*však* řadíme ke spojkám, ačkoli v textech převažuje užití částicové.

Diskusi uvádíme tabulkou 6.5.9, která udává normalizované ARF všech rozlišovaných slovních druhů (tedy včetně interpunkce a tvarů nerozpoznaných morfologickou analýzou) ve třech vybraných subkorpusech: pub_1992, pub_2000 a pub_2009. Údaje vznikly zadáním dotazu na slovní druh (např. tedy `[tag="N.*"]` pro substantiva) a normalizací výsledné ARF v daném subkorpusu na 100 milionů pozic. Ve 3. a 5. sloupci je uveden procentuální posun mezi sousedními dvěma hodnotami, v posledním sloupci je celkový posun mezi pub_1992 a pub_2009. Vzhledem k povaze ARF nemá smysl její hodnoty sčítat, proto také nejsou sloupcové součty hodnot normalizované ARF v jednotlivých subkorpusech shodné.

Z tabulky je na první pohled patrný poměrně výrazný pokles jmen, tj. substantiv (obr. 6.5.1), adjektiv (obr. 6.5.3) a číslovek, který je kompenzován nárůstem sloves (obr. 6.5.4), adverbií a částic.[17] Interpretace tohoto posunu není jednoduchá, některé jeho příčiny budou naznačeny v podkapitole 6.6. S poklesem jmenných slovních druhů a nárůstem slovesných nekoresponduje pouze nárůst předložek, jejichž frekvence by měla podle očekávání klesat spolu se jmény. Tento nárůst se nepodařilo uspokojivě

[17] Čísla týkající se částic sice nejsou příliš spolehlivá, jejich nárůst je však podpořen také nárůstem u adverbií, s nimiž bývají částice nejčastěji zaměňovány.

slovní druh	pub_1992	posun	pub_2000	posun	pub_2009	celkem
substantiva	21 435 447	2 %	21 783 881	−6 %	20 575 121	**−4 %**
adjektiva	6 833 595	−2 %	6 693 244	−8 %	6 164 761	**−10 %**
zájmena	4 070 985	1 %	4 114 469	12 %	4 616 450	**13 %**
číslovky	1 674 268	1 %	1 688 318	−6 %	1 578 793	**−6 %**
slovesa	8 089 080	11 %	8 986 978	10 %	9 907 572	**22 %**
adverbia	2 724 370	5 %	2 861 102	13 %	3 232 103	**19 %**
předložky	6 437 648	5 %	6 776 664	2 %	6 881 338	**7 %**
spojky	3 633 725	−3 %	3 519 007	4 %	3 671 689	**1 %**
částice	608 224	4 %	630 107	9 %	688 069	**13 %**
citoslovce	6 497	−24 %	4 964	37 %	6 791	**5 %**
interpunkce	11 394 763	−3 %	11 060 678	0 %	11 065 838	**−3 %**
nerozpoznaný	500 168	−15 %	424 353	21 %	513 428	**3 %**

Tabulka 6.5.9: Normalizovaná ARF slovních druhů ve vybraných subkorpusech.

vysvětlit ani přepočítáním tabulky 6.5.9 z ARF na frekvence, částečnou odpověď naznačuje pouze obr. 6.5.2 s neklesajícím (dokonce mírně rostoucím) frekvenčním průběhem substantiv v předložkové vazbě. V případě zájmen (obr. 6.5.5) je významný především nárůst zájmen ukazovacích a osobních (obr. 6.5.6), konkrétně pak reflexivního *se*. U spojek k žádnému výraznějšímu posunu nedochází, čísla týkající se citoslovcí nebereme v úvahu vzhledem k jejich nízké frekvenci a velké oscilaci.

Vraťme se však ke konkrétním lemmatům. Většina substantiv vykazuje pokles, což se týká hlavně substantiv spojených s politikou a formálním vyjadřováním: $^{\mathrm{M}}$*člen*, $^{\mathrm{P}}$*jednání*, $^{\mathrm{P}}$*prostředek*, $^{\mathrm{P}}$*předseda*, $^{\mathrm{M}}$*republika*, $^{\mathrm{P}}$*vláda*, $^{\mathrm{M}}$*země*. Pozvolný nárůst (takto budeme nadále označovat nárůst, který v letech 1992–2009 dosahuje nejvýše 50 %) zaznamenala substantiva $^{\mathrm{MP}}$*místo* a $^{\mathrm{M}}$*rok*,[18] větší nárůst pak $^{\mathrm{M}}$*člověk* a $^{\mathrm{P}}$*mluvčí*. Za pozornost stojí zejména lemma $^{\mathrm{M}}$*člověk*, za jehož nárůstem stojí především supletivní plurál *lidé*. Tento nárůst má pravděpodobně více příčin, kromě odpolitizování publicistiky a jejího tematického přiblížení „lidem“ se jednou z nich zdá být i rostoucí obliba používání plurálu *lidé* jako nekonkrétního vyjádření subjektu nebo objektu ve výrazech typu „lidé z okolních vesnic“ nebo „lidé na výstavě“ namísto konkrétnějších (a také formálnějších) alternativ „obyvatelé okolních vesnic“ nebo „návštěvníci výstavy“.

[18] Včetně plurálových tvarů *léta*, *let*, *letům* atd. Ačkoliv existuje také samostatné lemma *léto*, frekvence lemmatu *rok* není chybami desambiguace příliš ovlivněna díky tomu, že celková frekvence všech plurálových tvarů s dlouhým *é* je ve srovnání s ní zanedbatelná.

Z adjektiv zaznamenala pozvolný nárůst lemmata [P]*celý* a [P]*další*; pokles lemmat [P]*německý* a [M]*zahraniční* je výraznější a vzhledem k již zmiňovanému odklonu publicistiky od zahraničněpolitických témat není ani nijak překvapivý.

U zájmen [M]*co*, [M]*on*,[19] [MP]*oni*, [MP]*se*,[20] [M]*ten* převažuje nárůst, který je však pouze pozvolný. Nárůst je způsoben zejména vzrůstající neformálností publicistiky a také množstvím rozhovorů a přímé řeči v ní; tento druhý důvod je významný zvláště u osobních zájmen (obr. 6.5.6). Poznamenejme také, že zájmeno [M]*ten* je nefrekventovanějším demonstrativem, z jeho tvarů je pak dominantní tvar *to*, který však má samostatné (částicové) lemma pouze výjimečně. Důsledkem je, že téměř polovinu všech výsledků dotazu na ukazovací zájmena `[tag="PD.*"]` tvoří právě tvar *to* a že právě vzrůstající frekvence tohoto tvaru je pro frekvenční nárůst demonstrativ rozhodující. Použitá lemmatizace a slovnědruhové značkování sice pod zájmeno [M]*ten* zahrnuje také některé výskyty tvaru *to*, které by měly být označeny jako samostatná částice, těchto chyb je ale relativně málo a na zmiňovaný nárůst demonstrativ nemají výrazný vliv.

Naopak pokles zaznamenalo posesivum [P]*jeho* (a také posesiva jako celek) a demonstrativum [P]*tento*. To je pravděpodobně nahrazováno méně formálním [M]*ten*, což naznačuje také procentuálně podobný nárůst, který zaznamenalo zájmeno *tenhle*; jeho frekvence je však stále zhruba o řád menší než frekvence zájmena [P]*tento*.

Frekvenční nárůst zaznamenaly prakticky všechny druhy číslovek psaných slovy (obr. 6.5.7) včetně základních. Číslovku [P]*tři* tedy nacházíme ve výsledcích zřejmě proto, že její pozvolný nárůst byl vyhodnocen jako nejpravidelnější a s dostatečnou frekvencí. Nárůst lemmatu [M]*pár* je výraznější a souvisí se vzrůstající neurčitostí některých výrazů, což ilustruje výrazný nárůst kombinací [MP]*pár den* a [M]*pár minuta* (obr. 6.2.1 na straně 96). Podobný jev popisuje Millar (2009, str. 214): „It may be that readers of TIME have grown to prefer more speculation about what *may* or *could* be behind the stories in the news, or what is *likely* to happen in the future.“ Tomu kromě mírně se zvyšující frekvence modálních sloves odpovídá také nárůst lemmat *asi*, *možná*[21] a [M]*třeba*, kterým se budeme zabývat dále v části věnované adverbiím.

Zajímavostí je frekvenční průběh číslovky *čtyři* (obr. 6.5.9) vykazující známky periodicity podobné „volebním“ výrazům v podkapitole 6.3. Tato periodicita vynikne u kombinace *čtyři rok* (obr. 6.5.10), která se vztahuje typicky na délku funkčního období, a potvrzuje tak ovlivnění frekvenčního průběhu číslovky *čtyři* zdánlivě nesouvisejícím tématem.

Jedinou, ale velice početnou výjimkou ze všeobecného nárůstu číslovek je pokles frekvence číslovek psaných čísly (obr. 6.5.8). Tento pokles je pochopitelně ovlivněn použitými čistícími metodami a složením korpusu (např. televizními programy ve

[19] Zahrnuje také tvary zájmen *ona* a *ono*.
[20] Zvratné *se* je lemmatizováno vždy samostatně, nikdy tedy není součástí reflexiva.
[21] Chyby lemmatizace způsobené homonymií s adjektivem *možný* nejsou významné.

starších ročnících MFD), jak ale vyplývá z tabulky 6.5.9, celkový podíl číslovek tak mírně klesá.

Ve výsledných tabulkách najdeme celkem 21 sloves, všechna přitom vykazují frekvenční nárůst: $^{\mathrm{M}}$*být*,[22] $^{\mathrm{MP}}$*čekat*, $^{\mathrm{M}}$*dát*, $^{\mathrm{M}}$*dělat*, $^{\mathrm{P}}$*dokázat*, $^{\mathrm{MP}}$*dostat*, $^{\mathrm{MP}}$*chtít*, $^{\mathrm{M}}$*chystat*, $^{\mathrm{P}}$*mít*, $^{\mathrm{P}}$*moci*, $^{\mathrm{P}}$*muset*, $^{\mathrm{MP}}$*najít*, $^{\mathrm{P}}$*nechat*, $^{\mathrm{P}}$*plánovat*, $^{\mathrm{MP}}$*pomoci*, $^{\mathrm{MP}}$*přijít*, $^{\mathrm{M}}$*říkat*, $^{\mathrm{M}}$*snažit*, $^{\mathrm{M}}$*udělat*, $^{\mathrm{M}}$*vědět*, $^{\mathrm{MP}}$*začít* (obr. 6.4.7 na straně 135). Zároveň jsou také velice frekventovaná, jak ukazuje srovnání s FDC: pouhá tři z nich ($^{\mathrm{M}}$*chystat*, $^{\mathrm{P}}$*plánovat*, $^{\mathrm{MP}}$*pomoci*) podle něj nepatří mezi 50 nejfrekventovanějších českých sloves, naopak z první desítky sloves podle FDC chybějí ve výsledných tabulkách pouze dvě (*jít*, *říct*), jejichž frekvenční průběh sice není tak pravidelný, přesto vykazují celkový nárůst. Pro vysvětlení těchto zjištění je podstatný frekvenční nárůst sloves jako celku (viz tabulka 6.5.9), který činí 22 %. Ten sám o sobě vysvětluje pozvolný nárůst nejfrekventovanějších z nich, tedy zejména pomocných a modálních sloves $^{\mathrm{M}}$*být*, $^{\mathrm{P}}$*mít*, $^{\mathrm{P}}$*moci*, $^{\mathrm{P}}$*muset*, který se okolo této hodnoty pohybuje. Frekvenční nárůst většiny ostatních sloves je však výraznější, protože se přidávají další příčiny popisované v podkapitole 6.4.

K adverbiím připojujeme také jediné lemma z výsledných tabulek, které má většinu výskytů ve zdrojových datech označenu jako částice ($^{\mathrm{P}}$*nakonec*); toto označení je však podle našeho názoru sporné, také podle FDC převažuje adverbiální užití tohoto lemmatu. Protože existují i opačné případy označené v datech jako adverbia, u nichž podle FDC převažuje naopak užití v platnosti částice ($^{\mathrm{MP}}$*hlavně*, $^{\mathrm{P}}$*navíc*, $^{\mathrm{MP}}$*také*, $^{\mathrm{M}}$*třeba*), oba slovní druhy v diskusi spojíme.

Ve výsledných tabulkách najdeme celkem 15 adverbií a částic: $^{\mathrm{MP}}$*hlavně*, $^{\mathrm{M}}$*hned*, $^{\mathrm{MP}}$*hodně* (obr. 6.4.1 na straně 132), $^{\mathrm{P}}$*ještě*, $^{\mathrm{MP}}$*kde*, $^{\mathrm{P}}$*nakonec*, $^{\mathrm{P}}$*navíc*, $^{\mathrm{MP}}$*pak*, $^{\mathrm{MP}}$*tak*, $^{\mathrm{MP}}$*také*, $^{\mathrm{M}}$*tam*, $^{\mathrm{M}}$*teď*, $^{\mathrm{M}}$*třeba*, $^{\mathrm{MP}}$*už* (obr. 6.2.2 na straně 96), $^{\mathrm{M}}$*zase*. Podobně jako u sloves platí, že jde o velmi frekventovaná lemmata: vytvoříme-li na základě FDC seznam 50 nejfrekventovanějších lemmat označených jako adverbium nebo částice, nenajdeme v něm pouze čtyři z nich ($^{\mathrm{MP}}$*hlavně*, $^{\mathrm{M}}$*hned*, $^{\mathrm{P}}$*nakonec*, $^{\mathrm{P}}$*navíc*). Toto číslo je větší než u sloves hlavně proto, že se na předních místech seznamu vyskytuje velké množství adverbií a částic typických pro neformální mluvené projevy (*no*, *jo*, *prostě*, *teda* atd.), jejichž frekvence v psané publicistice je mnohem nižší.

Také všechna adverbia a částice vykazují nárůst, což je opět způsobeno především tím, že podle tabulky 6.5.9 činí nárůst adverbií 19 % a částic 13 %; zhruba tuto hodnotu také vykazuje nárůst lemmat $^{\mathrm{P}}$*ještě*, $^{\mathrm{MP}}$*kde*, $^{\mathrm{P}}$*nakonec*, $^{\mathrm{MP}}$*tak*. U ostatních lemmat je ještě výraznější a má v zásadě dvě příčiny. Jednou z nich je opět vzrůstající neformálnost publicistiky: lemmata $^{\mathrm{M}}$*tam*, $^{\mathrm{M}}$*teď*, $^{\mathrm{M}}$*zase* jsou v FDC označena značkou $-P$ jako netypická pro odbornou literaturu (a tedy pro formální vyjadřování), nárůst $^{\mathrm{MP}}$*hlavně* ve zdrojových subkorpusech koresponduje s poklesem *především*, s nímž je ve většině případů zaměnitelné. Druhým důvodem je rostoucí podíl užití těchto lemmat

[22] Včetně výskytů všech tvarů *být* jako pomocného slovesa.

v částicové platnosti, který textům dodává rysy mluvenosti (to potvrzuje výskyt řady z nich v rozhovorech) a který ovšem nemusí být vyvážen poklesem užití v platnosti adverbiální. To je také případ lemmatu [M]*hned* (obr. 6.5.11), v rámci jehož zhruba dvojnásobného frekvenčního nárůstu je patrný ještě výraznější nárůst jeho částicového užití, jak na kombinaci [P]*hned několik* ukazuje obr. 6.5.12. Dalším příkladem může být lemma [M]*třeba* (obr. 6.5.13), jeho celková frekvence vzrostla mezi pub_1992 a pub_2009 zhruba o třetinu, zatímco frekvence jeho adverbiálních užití v přísudku poklesla, jak je vidět na obr. 6.5.14 (použitý dotaz sice nenajde všechny takové výskyty, podstatné však je, že kromě nich nenajde prakticky žádné jiné).

Z předložek ve výsledných tabulkách zaznamenaly výrazný nárůst pouze nepůvodní předložky [MP]*díky*[23] a [M]*kvůli*. Protože mají obě podobný význam a rozdíl mezi nimi se navíc pozvolna stírá (spojování [MP]*díky* s negativně hodnocenými příčinami), nelze jejich nárůst vysvětlit zaměňováním jedné za druhou. Ostatní předložky zaznamenaly buď pozvolný nárůst: [MP]*do*, [MP]*na* (obr. 6.5.15), [MP]*od*, [P]*po*, [P]*před*, [P]*přes*, nebo pokles: [P]*k*, [P]*mezi*, [P]*o*. Vysvětlení příčin těchto posunů není snadné a vyžaduje podrobnější analýzu, která však již přesahuje rámec této práce.

U spojek je situace poněkud odlišná, přesto však výrazný nárůst spojky [MP]*ale* (obr. 6.4.3 na straně 133) nelze vysvětlit pouhým poklesem potenciálních alternativ *avšak*, *ovšem*, [MP]*však* podobně jako pozvolný nárůst spojky [MP]*i*. Samostatnou studii zaslouží také podrobné vyhodnocení vývojových tendencí a způsobů užití spojek uvozujících podmínkové věty, které uvádíme v tabulce 6.5.10 i s jejich normalizovanou ARF ve vybraných subkorpusech (dotazy byly zadány vždy na lemma a bez ohledu na slovní druh, např. tedy `[lemma="jestli"]`). Ve výsledných tabulkách z nich najdeme [M]*jestli* (obr. 6.5.16) a [M]*když*, příklonka *-li* byla již jako frekventované lemma vykazující pravidelný pokles zmíněna v podkapitole 6.4.

Ve 2. skupině výrazů se nacházejí lexikální kombinace, které tvoří jmennou frázi, případně jsou její součástí. Dalším charakteristickým rysem těchto kombinací je to, že řada z nich již byla zmiňována v podkapitole 6.4 věnované vyhodnocení *tau*: [M]*bytový dům*, [MP]*celý záležitost*, [M]*doprovodný program*, [M]*fond národní*, [M]*liberecký kraj*, [M]*milión marka*, [MP]*národní majetek*, [MP]*ředitelství silnice*, [MP]*sdělovací prostředek*, [MP]*stránka www*, [M]*tisíc marka*, [MP]*webový stránka*, [M]*zahraniční firma*, [P]*základní škola*, [MP]*zlínský kraj*. Protože by další rozbor zmíněných kombinací nic nového nepřinesl, nebudeme se jimi už zabývat.

Kombinace 2. skupiny můžeme podobně jako v podkapitole 6.4 rozdělit na dvě podskupiny podle toho, zda byl zaznamenán nárůst nebo pokles. I jejich příčiny jsou podobné, frekvenční nárůst je odrazem nových technologií ([M]*internetový stránka*), vzniku krajů ([M]*olomoucký kraj*, [M]*pardubický kraj*), rozvoje občanské společnosti

[23] Jde výhradně o předložku, od substantiva *dík* je odlišena zněním lemmatu; existuje však také samostatné předložkové lemma *dík*.

spojka	pub_1992	posun	pub_2000	posun	pub_2009	celkem
jestli	6 454	21 %	7 841	66 %	12 989	**101 %**
jestliže	6 766	−50 %	3 363	−28 %	2 414	**−64 %**
kdyby	19 312	−20 %	15 519	14 %	17 760	**−8 %**
když	69 644	11 %	77 059	23 %	94 454	**36 %**
-li	24 108	−59 %	9 788	−36 %	6 301	**−74 %**
pokud	33 051	3 %	33 994	14 %	38 636	**17 %**
zda	19 111	−6 %	17 936	−11 %	15 875	**−17 %**
zdali	563	−47 %	300	−16 %	253	**−55 %**

Tabulka 6.5.10: Normalizovaná ARF některých spojek ve vybraných subkorpusech.

([M]*občanský sdružení*) a množství sportovního zpravodajství ([P]*tři bod*), projevuje se také měnící se zaměření publicistiky na dopravu, výstavbu, zdravotnictví, školství a podobná prakticky orientovaná témata ([P]*centrum město*, [P]*metr čtvereční*, [MP]*rodinný dům*, [M]*stavební povolení*, [MP]*střední škola*, [P]*tisíc koruna*, [P]*záchranný služba*). Kombinace [P]*policejní mluvčí* typicky doprovází přímou řeč a kombinace [MP]*pár den*, [M]*pár minuta* (obr. 6.2.1 na straně 96) jsou zřejmě projevem vzrůstající neurčitosti vyjadřování v publicistice. Tuto hypotézu potvrzuje také nárůst kombinace *pár hodina*, na rozdíl od pouze oscilujících *několik den*, *několik hodina* a pozvolna a nepravidelně rostoucí *několik minuta*.

Pokles zaznamenaly především kombinace odrážející odklon publicistiky od původní převážně politické orientace: [MP]*akciový společnost*, [M]*ministerstvo vnitro*, [MP]*ministr vnitro*, [MP]*ministr zahraničí*, [P]*předseda vláda*, [M]*spojený stát* (Spojené státy), [P]*státní rozpočet*, [P]*státní správa*, [P]*tiskový konference*, [M]*valný hromada*. Jde o další potvrzení již několikrát zmíněného trendu, stejně jako pokles kombinací spíše formálního charakteru: [M]*oba strana*, [P]*podstatný část*, [P]*značný část*; ke druhým dvěma poznamenejme, že výrazný pokles vykazují i samostatná lemmata *podstatný* a *značný*. Jistou zvláštností je kombinace [M]*minulý týden*, která sice zaznamenala v MFD pokles, v publicistice však spíše mírný nárůst, takže její frekvenční průběh hodnotíme celkově spíše jako oscilaci.

Do 3. skupiny byly zařazeny lexikální kombinace, které netvoří jmennou frázi. Na rozdíl od kombinací 2. skupiny odráží většina z nich posuny v jazyce publicistiky, nikoli pouze její tematickou orientaci nebo měnící se realitu. Pouze tři tyto kombinace se již objevily v podkapitole 6.4 věnované vyhodnocení *tau* ([MP]*tento souvislost*, [M]*úplně jiný*, [M]*vysvětlit mluvčí*), většina tedy byla nalezena až *taumed* díky jejich vysoké frekvenci. Ve vztahu ke zdrojovým datům poznamenáváme, že nuly ve výsledných tabulkách MFD za rok 1995 jsou u kombinací [M]*doprovodný program*, [M]*vysvětlit mluvčí* pravděpodobně způsobeny fragmentárností tohoto ročníku (viz obr. 4.5.5 na straně 59).

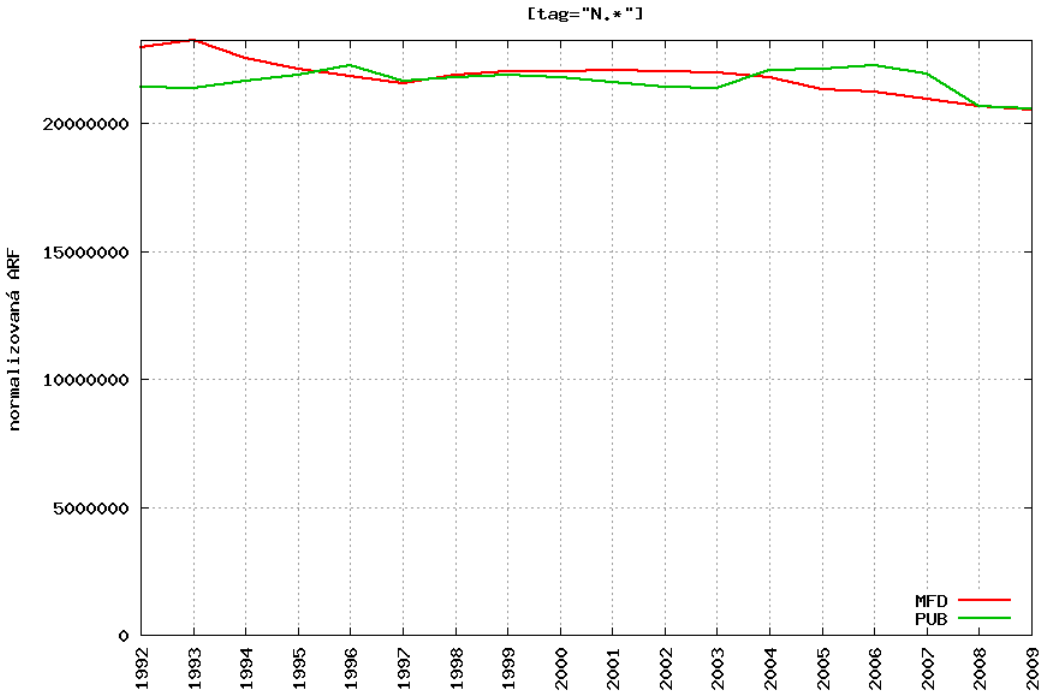
Mezi kombinacemi této skupiny najdeme několik spojovacích výrazů, jejichž znění ve výsledných tabulkách neobsahuje čárku, přestože v textu jsou oba členy čárkou typicky odděleny. To je dáno tím, že interpunkce byla při výpočtu ignorována (viz oddíl 5.4.2), takže například do kombinace [P] *doufat že* je zahrnuto také (a především) „doufat, že“. Ignorována však nebyla alfabetická lemmata, například zvratné *se*; kombinace [MP] *domnívat že* v sobě tedy nezahrnuje variantu se slovosledem „domnívat se, že“.

Ze spojovacích výrazů zaznamenaly nárůst [P] *doufat že*, [M] *myslet že*, [M] *vědět jestli*, [M] *věřit že* – ve všech případech přitom převažují tvary slovesa v 1. os. sg. a většina výskytů pochází z přímé řeči. Pokles naopak zaznamenaly formálnější kombinace [MP] *domnívat že*, [MP] *otázka zda* a také [P] *přesvědčení že* – ta se v publicistice na počátku 90. let používala typicky ve spojení „vyjádřil(a)/vyslovil(a) přesvědčení, že“, jehož frekvence je dnes již zanedbatelná (obr. 6.5.17). V současné publicistice je naopak běžné doslovné přebírání vyjádření tiskových mluvčích, jak dokumentuje nárůst kombinací [M] *říci mluvčí*, [MP] *uvést mluvčí*, [M] *vysvětlit mluvčí* a graf na obr. 6.5.18.

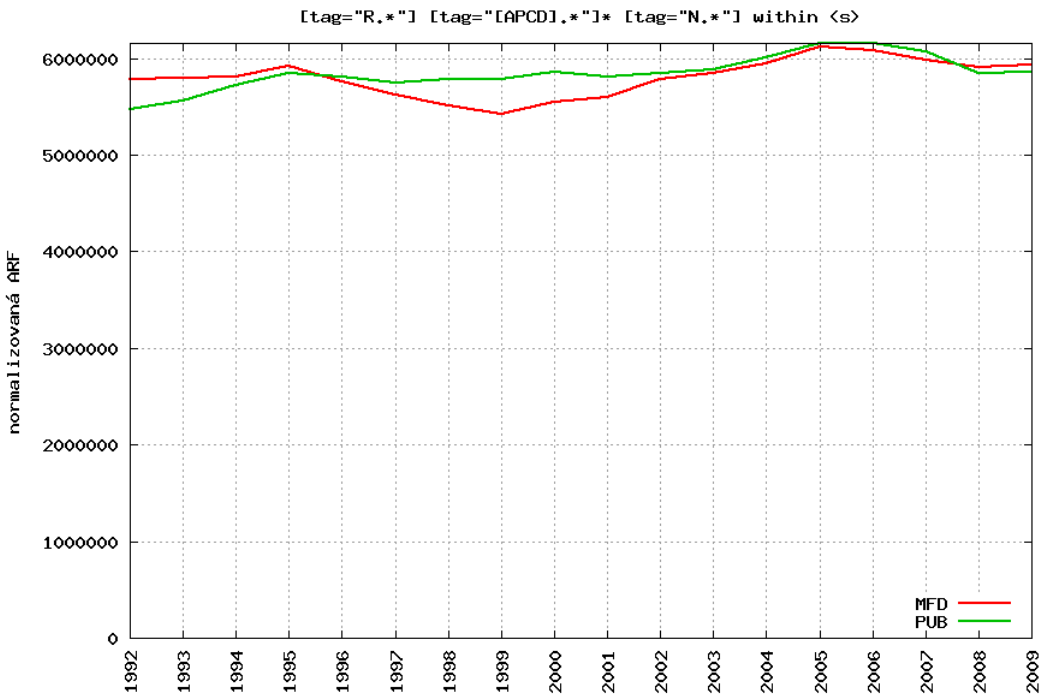
Dalším projevem posunu publicistiky směrem k mluvenému jazyku je frekvenční nárůst kombinací s modifikující částicí [P] *hned několik* (obr. 6.5.12), [M] *úplně jiný*, [M] *vůbec poprvé*. Nárůst zaznamenaly také kombinace, které lze charakterizovat jako obraty, které jsou stále populárnější při popisování konkrétních událostí nebo osob: [MP] *k vidění*, [P] *od začátek*, [P] *patřit mezi*, [M] *sloužit jako*. Převážně tématem jsou způsobeny frekvenční nárůsty kombinací [P] *o víkend*, [M] *začít stavět*.

Pokles zaznamenaly formální vyjadřovací prostředky, mezi nimiž najdeme především slovesné vazby [P] *dospět k*, [P] *označit za*, [MP] *považovat za*, [P] *přispět k*, [P] *směřovat k* a víceslovné předložky vyjadřující abstraktní vztahy: [P] *rozpor s* (v rozporu s), [P] *soulad s* (v souladu s), [P] *souvislost s* (v souvislosti s), [MP] *tento souvislost* (v této souvislosti), [P] *vzhledem k*; u víceslovných předložek [M] *po skončení*, [MP] *spolu s* je pokles pozvolnější. Frekvence poklesla také u kombinací [P] *spor mezi*, [MP] *vztah mezi*, které začátkem 90. let popisují téměř výhradně vztahy mezi organizacemi, zatímco v současné publicistice mezi nimi přibývá vztahů a sporů osobních. Pokles vykazuje také kombinace [P] *žádný případ* (v žádném případě), v případě kombinace [M] *poté co* jde spíše o oscilaci.

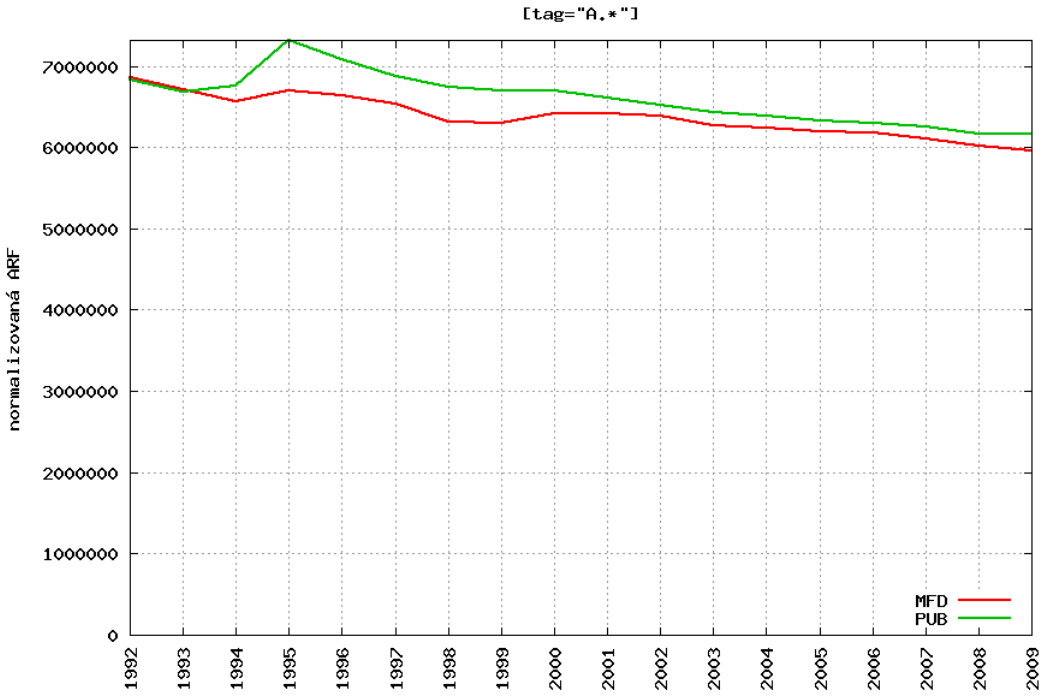
Hlavním přínosem této podkapitoly byla charakterizace jazykových změn u velmi frekventovaných výrazů. Byly zjištěny významné posuny ve frekvencích některých slovních druhů a také pozvolné frekvenční nárůsty u většiny frekventovaných slov z jádra slovní zásoby, které však nebyly uspokojivě vysvětleny. Na úrovni lexikálních kombinací byl zaznamenán frekvenční pokles formálních vyjadřovacích prostředků a naopak nárůst prostředků neformálních, který je důležitý pro dokreslení charakteru změn jazyka publicistiky.



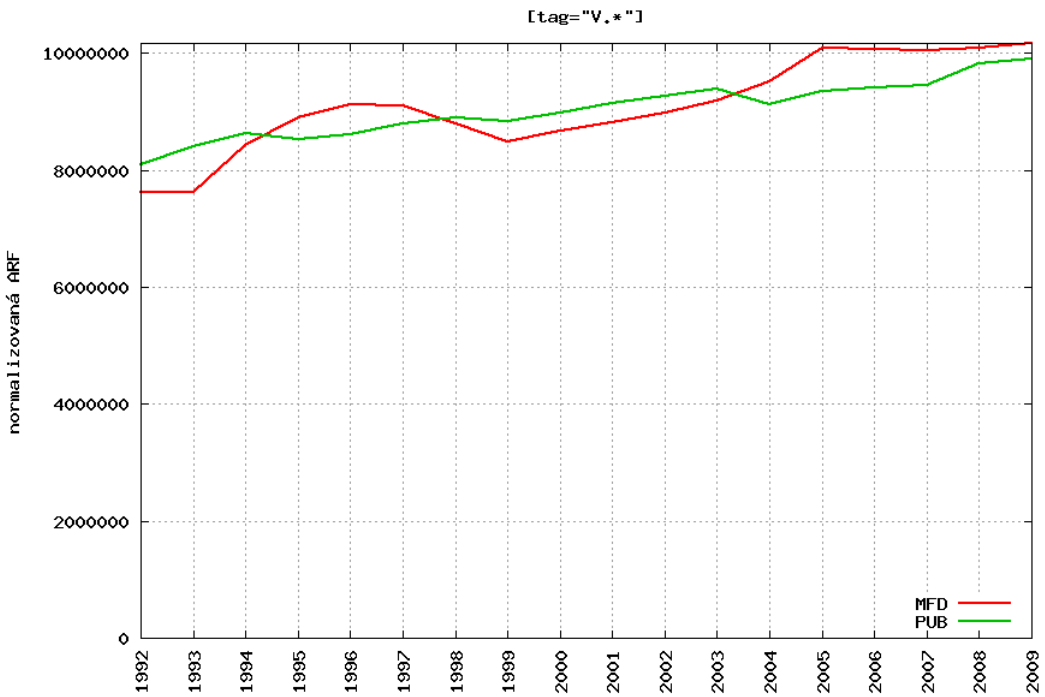
Obrázek 6.5.1: Průběh normalizované ARF substantiv.



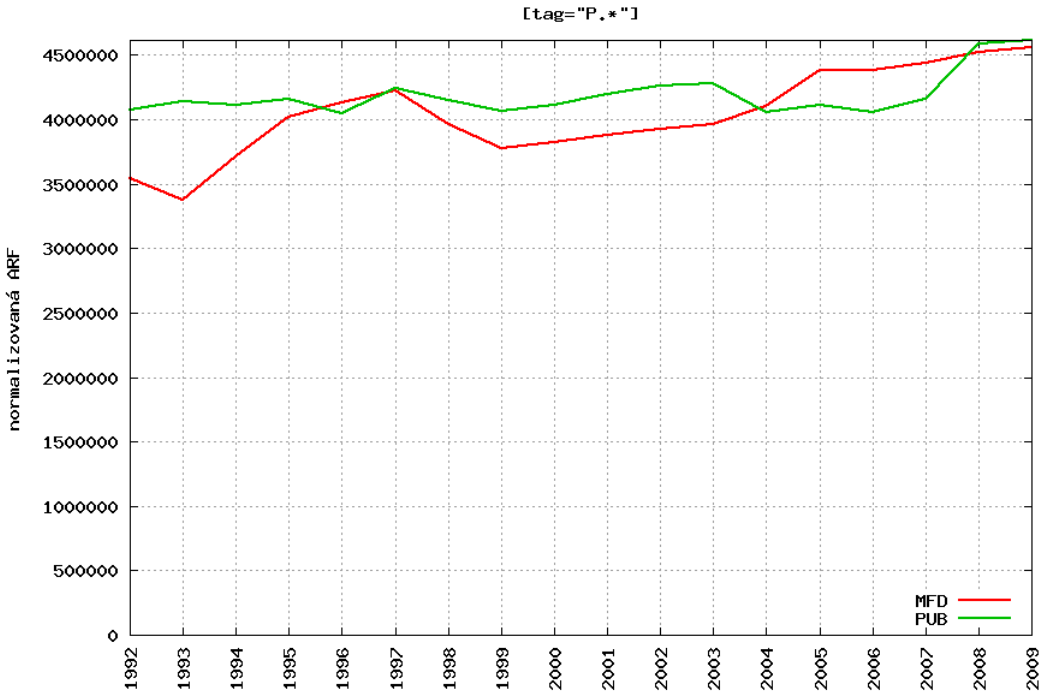
Obrázek 6.5.2: Průběh normalizované ARF substantiv v předložkové vazbě.



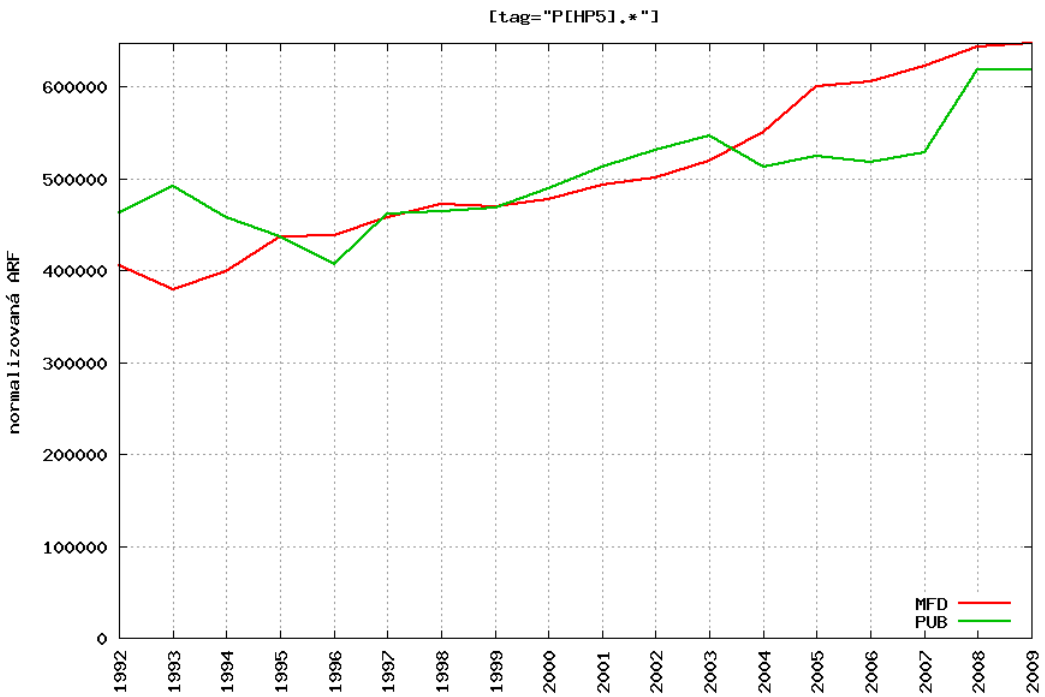
Obrázek 6.5.3: Průběh normalizované ARF adjektiv.



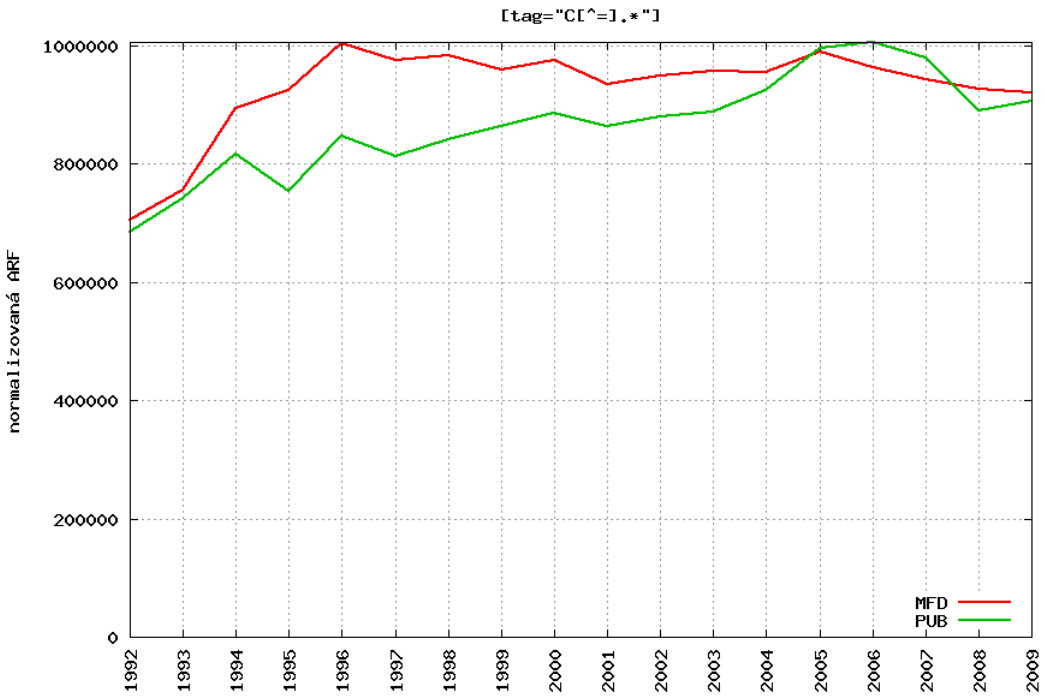
Obrázek 6.5.4: Průběh normalizované ARF sloves.



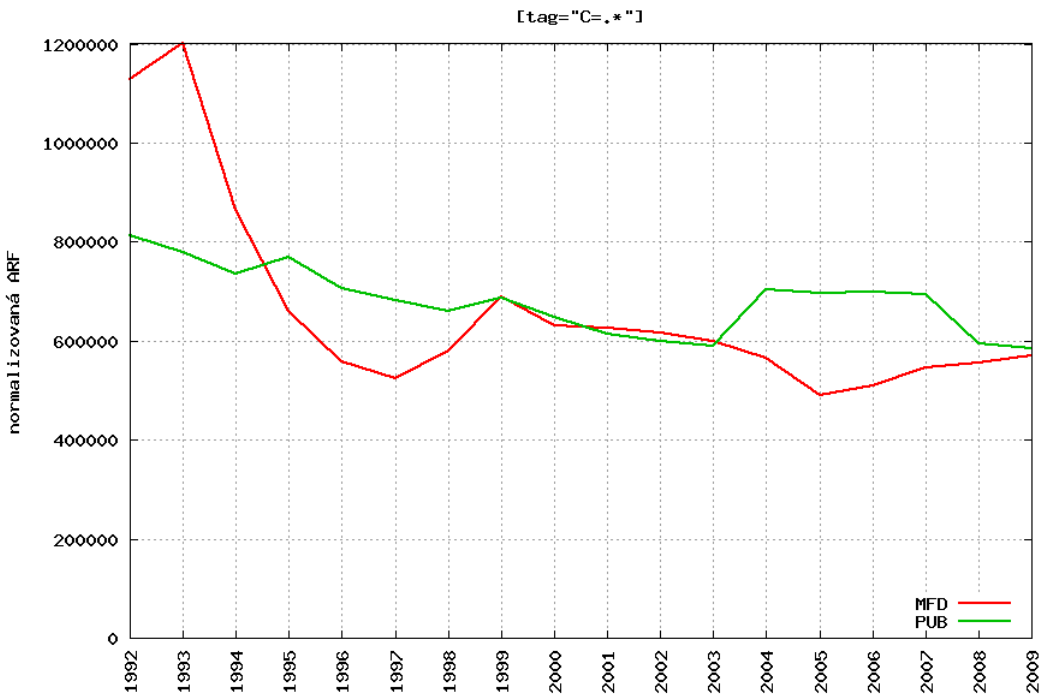
Obrázek 6.5.5: Průběh normalizované ARF zájmen.



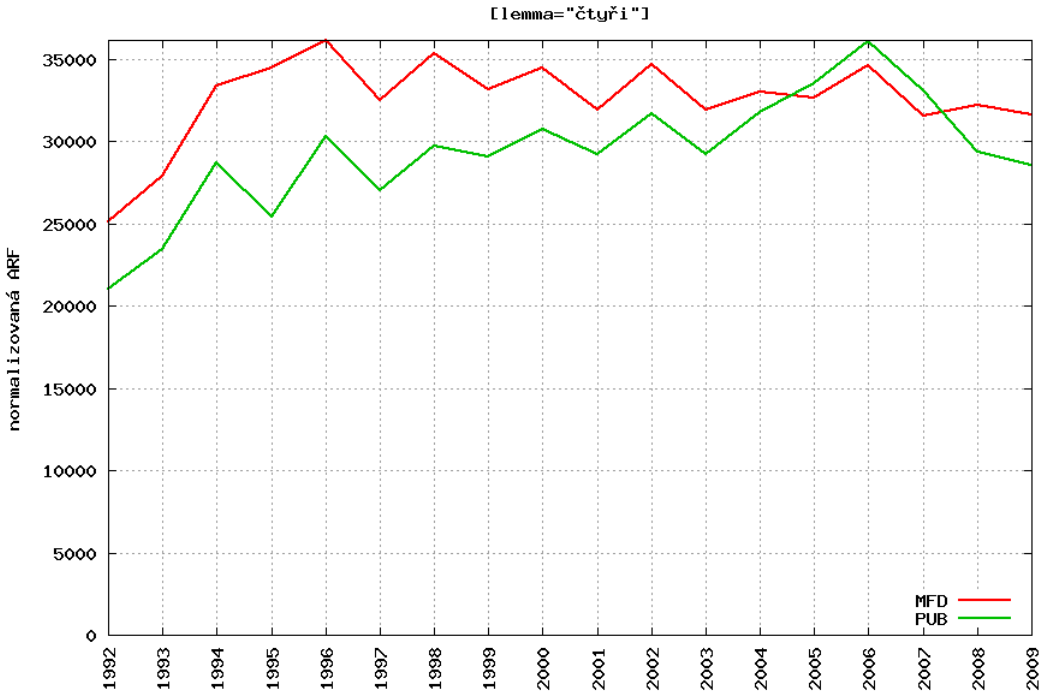
Obrázek 6.5.6: Průběh normalizované ARF osobních zájmen.



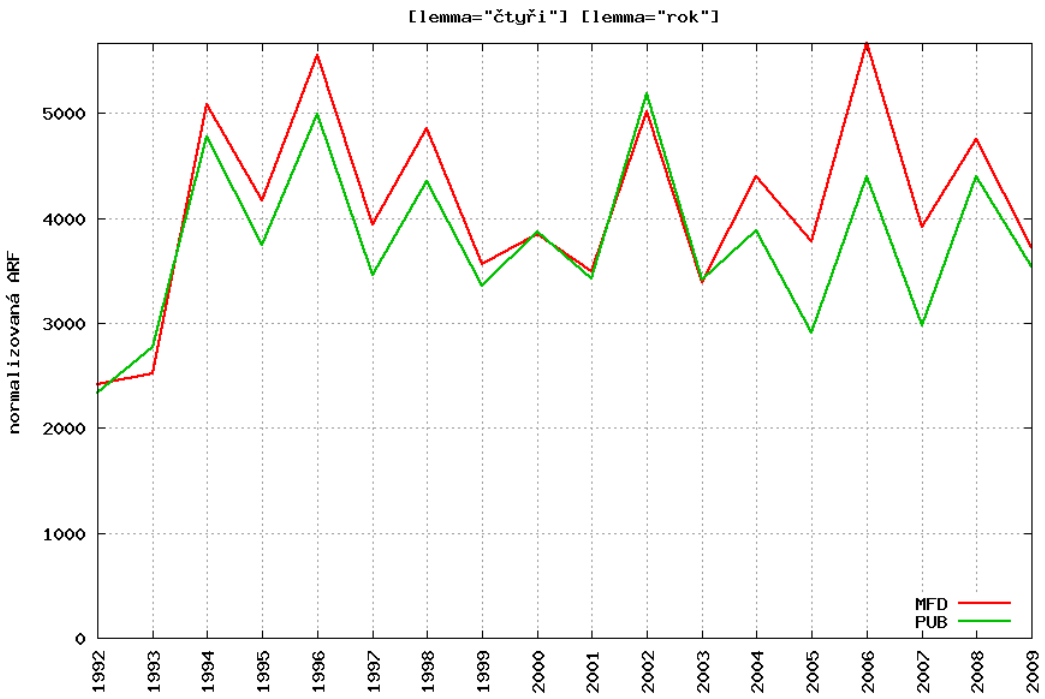
Obrázek 6.5.7: Průběh normalizované ARF číslovek psaných slovy.



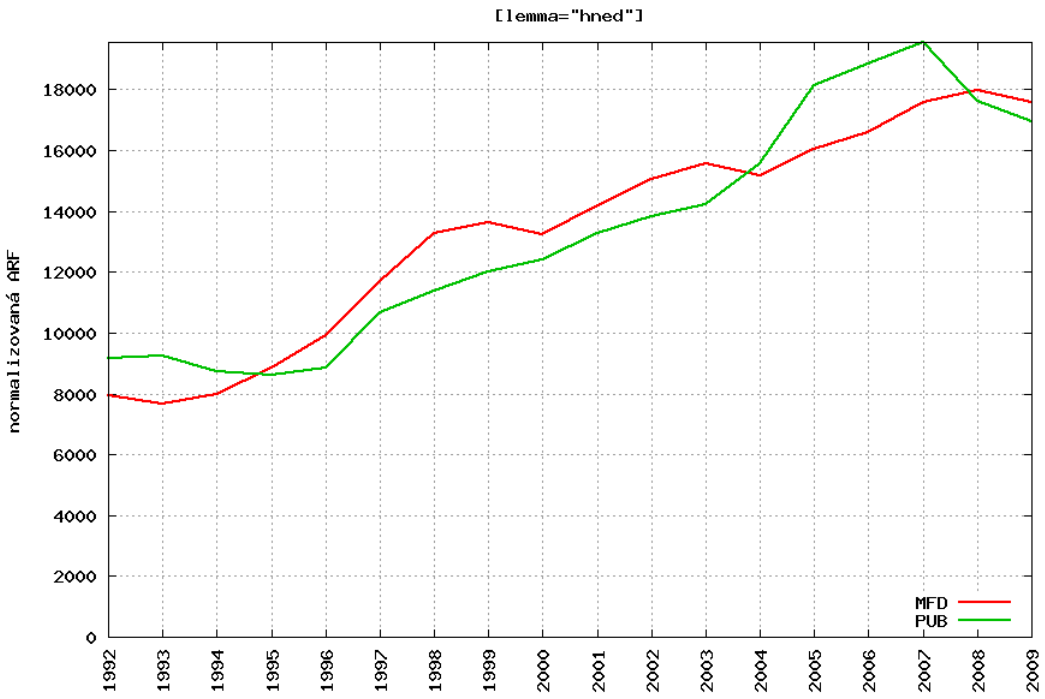
Obrázek 6.5.8: Průběh normalizované ARF číslovek psaných čísly.



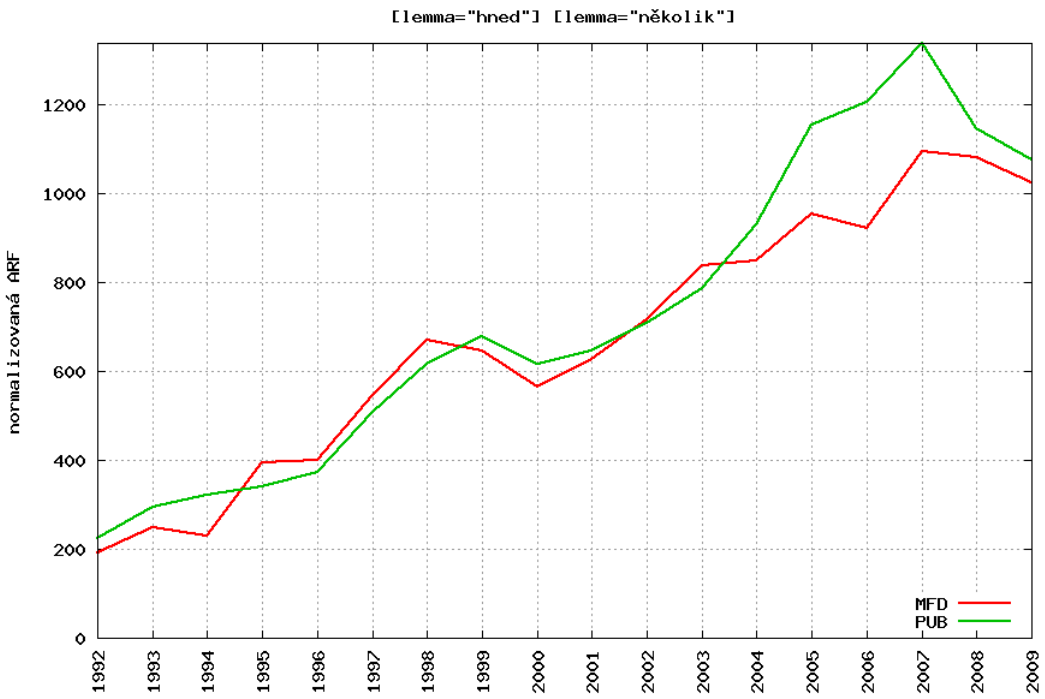
Obrázek 6.5.9: Průběh normalizované ARF lemmatu *čtyři*.



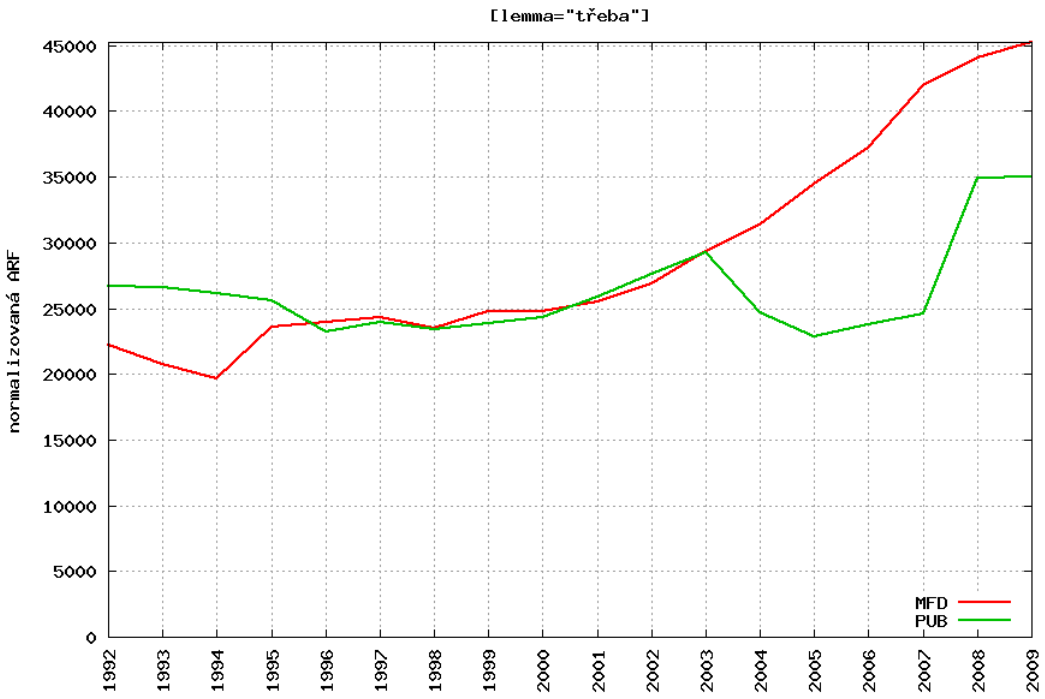
Obrázek 6.5.10: Průběh normalizované ARF kombinace *čtyři rok*.



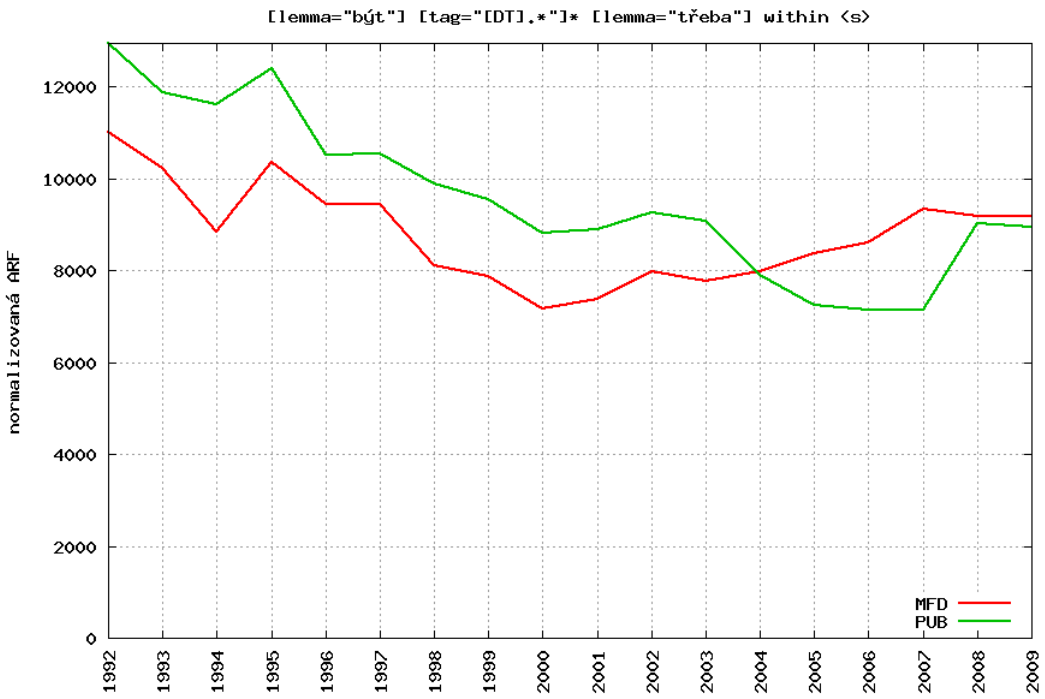
Obrázek 6.5.11: Průběh normalizované ARF lemmatu *hned*.



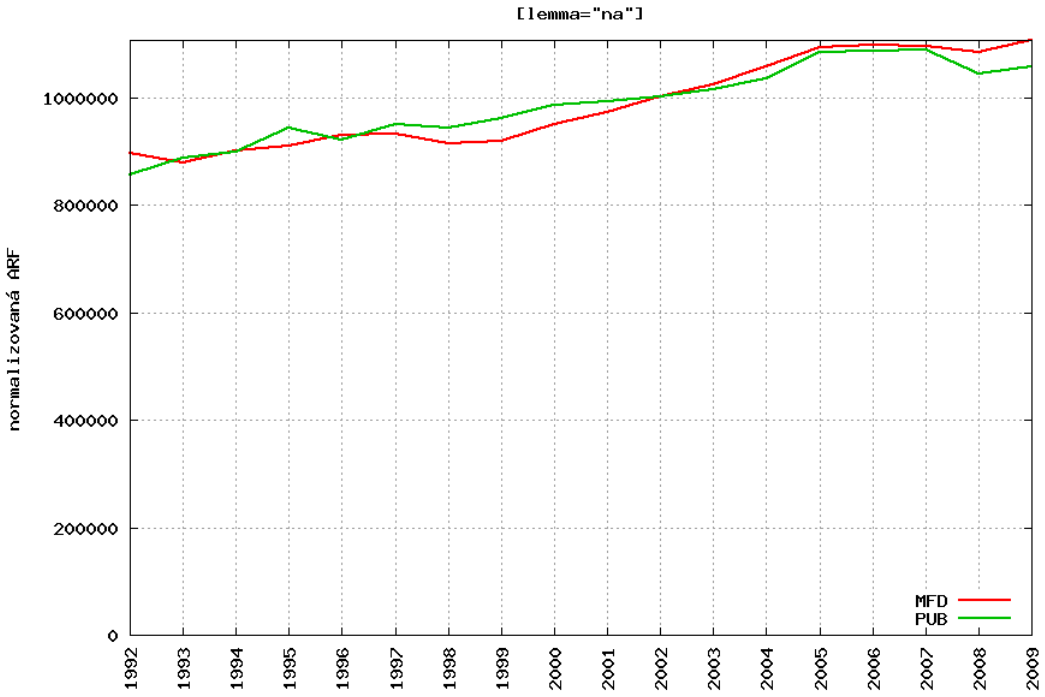
Obrázek 6.5.12: Průběh normalizované ARF kombinace *hned několik*.



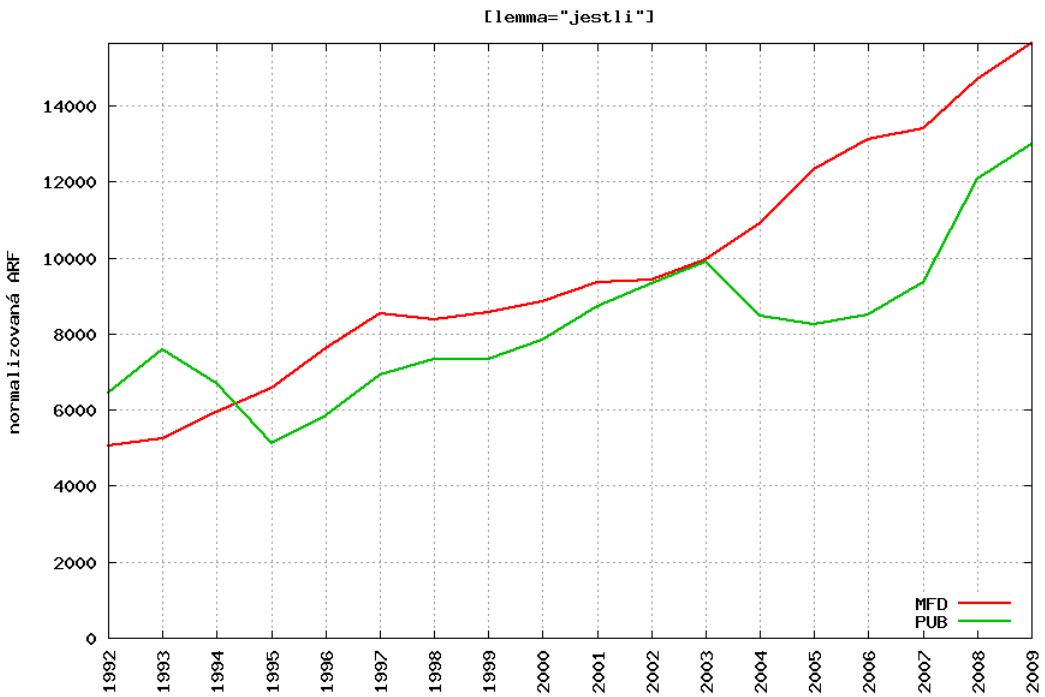
Obrázek 6.5.13: Průběh normalizované ARF lemmatu *třeba*.



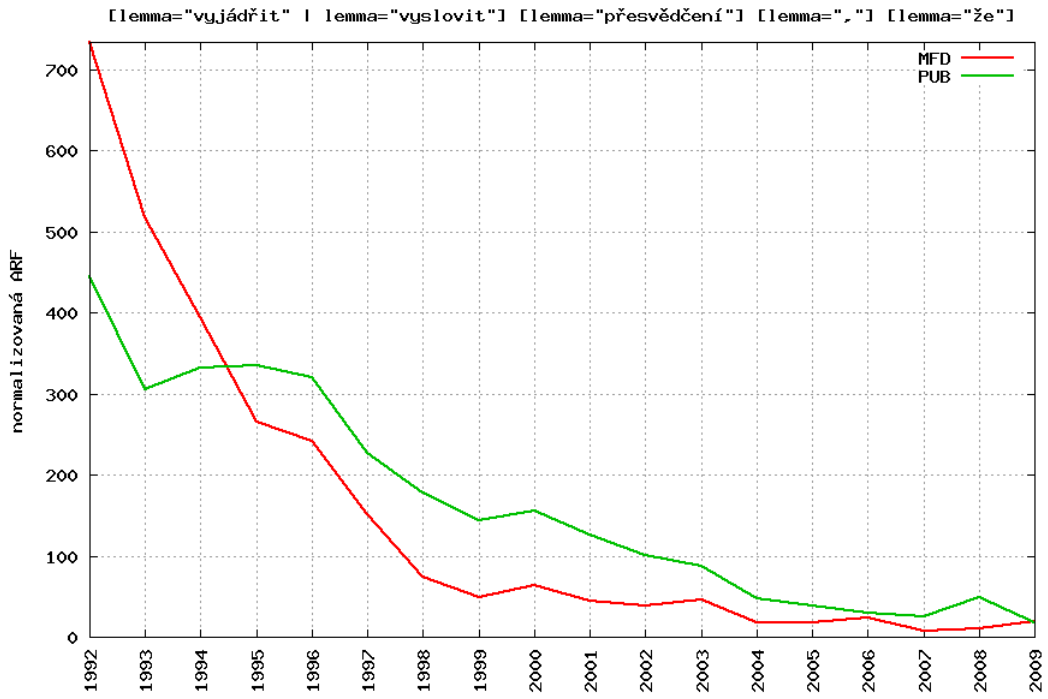
Obrázek 6.5.14: Průběh normalizované ARF lemmatu *třeba* v přísudku.



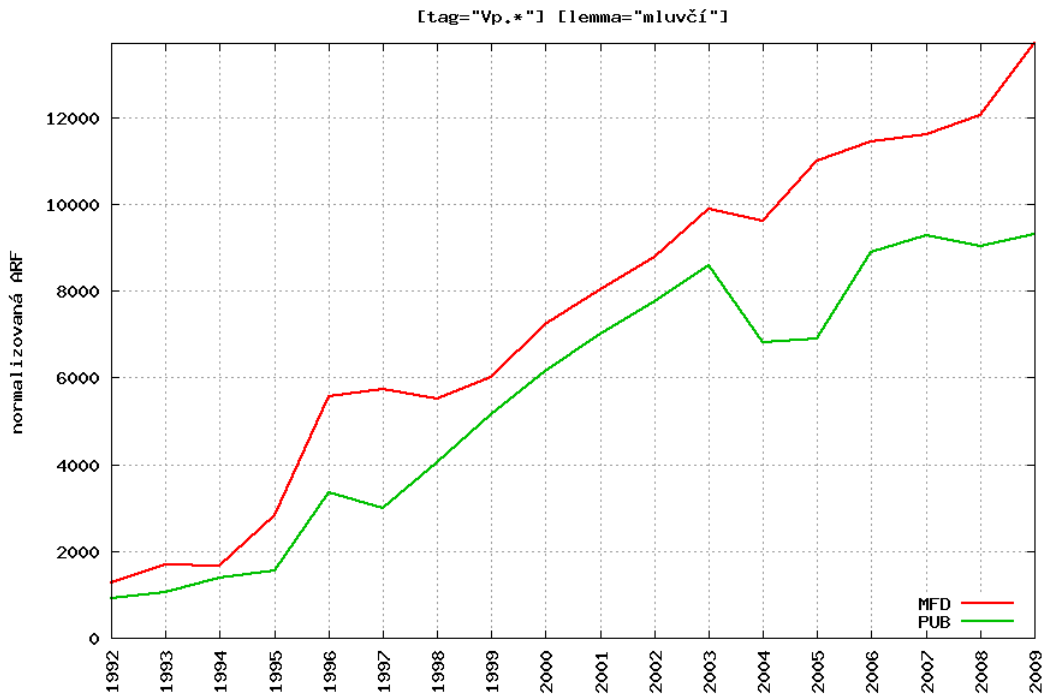
Obrázek 6.5.15: Průběh normalizované ARF lemmatu *na*.



Obrázek 6.5.16: Průběh normalizované ARF lemmatu *jestli*.



Obrázek 6.5.17: Průběh normalizované ARF spojení *vyjádřit/vyslovit přesvědčení, že*.



Obrázek 6.5.18: Průběh normalizované ARF spojení préterita s lemmatem *mluvčí*.

6.6 Shrnutí

V kapitole 6 jsme se podrobně zabývali výsledky tří metod aplikovaných na tři různé řady subkorpusů: reprezentativních repre_TTT_KKKK a publicistických pub_RRRR a mf_RRRR. V podkapitole 6.2 se však ukázalo, že pro diachronní srovnání blízkých stavů jazyka nejsou reprezentativní subkorpusy vhodné. Problémem je především nespolehlivost údajů vyvozovaných na základě malého množství časových bodů a relativně velká heterogenita dat spojená s proměnlivostí složení jednotlivých kategorií v čase. S tím souvisí sám koncept reprezentativnosti založený na recepci jazyka, který umožňuje (a teoreticky ospravedlňuje) zařazování textů, které vznikly i několik desítek let před daným obdobím, což obraz jazykových změn dále zamlžuje. Jejich studium by výrazně usnadnilo alespoň doplnění data vzniku textu do bibliografických údajů o jednotlivých textech, jehož současná absence dovršuje velice omezenou praktickou použitelnost textů (zvláště neperiodik) v reprezentativních korpusech pro detekci vývojových tendencí v jazyce.

Zmíněné negativní rysy reprezentativních korpusů jsou problematické především z hlediska prováděného diachronního srovnání a neznamenají nutně zpochybnění dat v nich pro řadu jiných účelů. I z mnoha dalších úhlů pohledu je však problematická kategorie „životní styl“ a její zařazení do odborné literatury; tato kategorie by měla být překlasifikována a její podstatná část přesunuta do publicistiky, v úvahu připadá i vznik nového hlavního typu textu jako obdoby „popular magazines“ v korpusu COCA (viz oddíl 3.3.5). To by sice vyžadovalo přepracování procentuálního zastoupení některých dalších kategorií, tomu se ale v budoucnu stejně nebude možné vyhnout už vzhledem k tomu, že zastoupení hlavních typů textu se mezi SYN2000 a SYN2005 změnilo příliš radikálně (viz oddíl 5.3.1) a shodné složení korpusů SYN2005 a SYN2010 je z hlediska recepce jazyka obtížně obhajitelné (viz podkapitola 3.2).

Pro diachronní srovnání blízkých stavů jazyka je proto vhodnější použít publicistické subkorpusy i přesto, že se tak výpovědní hodnota výsledků zužuje z psaného jazyka pouze na publicistiku. V dalším textu se proto soustředíme především na výsledky dosažené na publicistických subkorpusech řady pub_RRRR a mf_RRRR.

Každá z metod aplikovaných na publicistické subkorpusy dává jiný druh výsledků: iterativní *cbf*, *chi*, *ll* vyzdvihují výraznou oscilaci, *tau* pravidelný vývojový trend a *taumed* výrazy s vysokou frekvencí. Tyto metody jsou výsledkem hledání vhodného způsobu srovnání jazykových dat, které jsou k dispozici v korpusech řady SYN, což se týká jak výběru metod, tak i subkorpusů, na něž byly aplikovány. Zvolené způsoby srovnání jsou tedy pouze jedněmi z mnoha, které byly vybrány vzhledem k různorodému charakteru výsledků popisovaných v jednotlivých podkapitolách. Některé z nich vypovídají o změnách v jazyce publicistiky, jiné pouze o proměnách doby

a jejích témat, řada jich také odkrývá nedostatky v anotaci a složení dat, které však nejsou tak podstatné jako v případě reprezentativních subkorpusů.

Chyby dané špatnou kategorizací publicistických textů, jejich nedostatečným čištěním nebo lemmatizací byly nalezeny zejména v podkapitole 6.3 založené na iterativních *cbf*, *chi*, *ll*. Těchto chyb je však relativně málo, zvláště vezmeme-li v úvahu, že tyto metody mají tendenci vyzdvihovat všechny nepravidelnosti. Chyby při čištění a kategorizaci textů se navíc týkají převážně starších textů.

Přes zmiňovanou různorodost použitých metod se však ve všech podkapitolách projevovalo problematické složení publicistiky. Publicistika tvoří v současné době nijak blíže nečleněnou kategorii 1. úrovně (tj. celý hlavní typ textu, viz také tabulka 4.3.1 na straně 44) v rozsahu přibližně jedné třetiny celého reprezentativního korpusu, v jejím složení tak dochází k velkým posunům (viz obr. 4.5.2 na straně 55), které mohou vést k rozkolísání frekvenčního průběhu pozorovaných výrazů. Složení publicistiky v reprezentativních korpusech by proto mělo být podrobněji stanoveno nejenom co do základních publicistických podtypů, zejména seriózní/bulvární a celostátní/regionální publicistiky, ale také jednotlivých významných titulů, jakými jsou MFD, LN, HN, Právo nebo Blesk. Nutným předpokladem pro stanovení těchto proporcí je rozlišování seriózní a bulvární publicistiky v anotaci, v současné době mezi nimi není rozdíl, takže například Respekt je stejně jako Blesk anotován značkami *txtype=PUB* a *genre=MIX*.

Nevhodným složením publicistiky se vyznačuje zejména korpus SYN, a v důsledku toho také publicistické subkorpusy řady pub_RRRR (viz obr. 4.5.4 na straně 58). Výsledkem je fakt, že frekvenční průběh řady lemmat a kombinací (včetně některých velmi frekventovaných) v nich lze rozdělit zhruba na tři období: 1992–1997 (rozkolísanost daná malým množstvím dat, nekompletními ročníky a nekonzistentním čištěním, která se týká také subkorpusů řady mf_RRRR), 1998–2003 (konzistentní složení dat, stabilní frekvenční průběhy) a 2004–2009 (velké výkyvy v subkorpusech řady pub_RRRR dané rozhodujícím podílem VLP hlavně v letech 2005–2007). Rozdělení na tato období výsledky sice neovlivňuje zásadně, negativní vliv složení dat je ale zřejmý zvláště v případě corpus-driven metod vyhledávajících pravidelný frekvenční průběh.

Ideálním řešením současného stavu složení publicistiky v korpusu SYN je doplnění dat. Chybějící publicistika ze začátku 90. let je pravděpodobně obtížně dostupná, řada novějších ročníků však dosud nebyla zveřejněna, přestože jsou již k dispozici v bance ČNK. Neméně důležité však je zavedení podrobnějšího členění publicistiky, které je zmíněno výše, a také vnitřního členění jednotlivých periodik podle sekcí či rubrik na jednotlivé tematicky zaměřené části, které by mělo být k dispozici alespoň u významnějších titulů. Obě tyto změny by zároveň usnadnily udržování srovnatelného složení publicistických subkorpusů.

Přes zmíněné nedostatky ve složení publicistických subkorpusů se však domníváme, že je možné z nich – při splnění určitých podmínek – vyvozovat spolehlivé závěry týkající se pozorovaných vývojových tendencí. Na základě zkušeností z této kapitoly bychom mohli jako v praxi postačující podmínku formulovat následující kritéria: dané lemma, kombinace nebo obecně jakýkoli jev by měl být frekventovaný a jeho vývojová tendence (nárůst nebo pokles) pravidelná a bez výkyvů.

Zdůrazněme, že pozorování by měla být založena na publicistických subkorpusech, v nichž může být zjištěná pravidelnost podpořena množstvím časových bodů, které by ve spojení s vysokou frekvencí měly prakticky vylučovat možnost ovlivnění výsledků složením dat. U reprezentativních korpusů prozatím nelze pravidelnost vývojových tendencí ověřit, k dispozici je příliš málo časových bodů v příliš variabilních datech.

Doposud zmiňované závěry byly spíše metodologické, zaměřené na složení dat a návrhy řešení zmiňovaných problémů. V jednotlivých podkapitolách se však objevila řada zjištění týkajících se změn v jazyce publicistiky, zejména její vzrůstající neformálnost. Studium její povahy a projevů od morfologie až po stylistiku vyžaduje podrobnější analýzu přesahující rámec této práce, v dalším textu proto pouze naznačíme, kudy by se mohla ubírat.

Podrobnější rozpracování by zasloužil především rozbor frekvenčních změn po slovních druzích, jejich podtypech a kombinacích (v širším smyslu, včetně kombinací více než dvouslovných), jakási obdoba *Statistik češtiny* (Bartoň et al., 2009) zachycující změny jazyka v čase. Vyjít lze také z tabulky 6.6.1 shrnující výsledky dotazu na pád substantiva, například tedy `[tag="N...1.*"]` pro nominativ. Tabulka ukazuje poměrně výrazný nárůst akuzativu (obr. 6.6.1), u ostatních pádů průběh buď kolísá, nebo je příliš pozvolný (instrumentál na obr. 6.6.2). Vyvozovat z tohoto pozorování obecnější závěry však považujeme za předčasné a málo podložené, potřebná by byla zejména podrobná charakterizace posunů v rámci jednotlivých pádových kategorií a jejich vztahu k tématu a stylu textu.

pád	pub_1992	posun	pub_2000	posun	pub_2009	celkem
Nom	5 223 757	4 %	5 453 698	−6 %	5 124 828	**−2 %**
Gen	4 827 816	5 %	5 082 289	−9 %	4 628 798	**−4 %**
Dat	562 288	1 %	565 326	−2 %	554 328	**−1 %**
Acc	2 922 907	11 %	3 258 756	6 %	3 444 215	**18 %**
Voc	12 390	−38 %	7 639	18 %	9 021	**−27 %**
Loc	2 062 686	5 %	2 165 392	−1 %	2 140 273	**4 %**
Ins	1 195 614	−1 %	1 180 079	−3 %	1 144 619	**−4 %**

Tabulka 6.6.1: Normalizovaná ARF pádů substantiv ve vybraných subkorpusech.

Budeme-li vycházet z tabulky procentuálního zastoupení slovních druhů v hlavních typech textu ve *Statistikách češtiny* (Bartoň et al., 2009, str. 130), zjistíme, že celkové posuny jednotlivých slovních druhů uváděné v tabulce 6.5.9 na straně 148 odpovídají posunu publicistiky směrem od odborné literatury k beletrii. Tento posun byl zjištěn u všech slovních druhů s výjimkou předložek a také spojek, u nichž dochází spíše k oscilaci, a potvrzuje tak pozorování, které zmiňují už Křen a Hlaváčová (2008, str. 445). Bylo by ovšem zjednodušující tvrdit, že se publicistika „beletrizuje“, některá zjištění tomu ostatně ani neodpovídají (např. vzrůstající podíl sloves v kategoriálním užití nebo frekvenční nárůst číslovek psaných slovy zmiňovaný v podkapitole 6.5). Je však pravděpodobné, že prodělává posun, kterým do ní proniká řada jevů pro beletrii typických, což koresponduje se vzrůstajícím podílem přímé řeči v rozhovorech.

Tento posun se nemusí týkat celé publicistiky, je právě tak možné, že se týká jenom některého druhu periodik (např. bulvárních, případně regionálních), zatímco jiných se týká jen okrajově (seriozní deníky a týdeníky) nebo je soustředěn pouze do některých jejich částí (zájmové přílohy). Vývoj, kterým česká publicistika od roku 1992 prošla, lze charakterizovat i jako diverzifikaci a jasnou profilaci jak na úrovni jednotlivých periodik, tak uvnitř nich. Tento vývoj však je bohužel za dosavadního stavu anotace publicistických textů obtížné kvantifikovat.

Na morfologické úrovni je možné analyzovat pronikání některých tvarů z obecné češtiny, a tedy jejich (v různé míře) vzrůstající přijatelnost pro psaný text. V mnoha případech se však bohužel nelze opřít o lemmatizaci a morfologické značkování, je tedy nutné vycházet ze slovních tvarů a tyto tvary dále ručně filtrovat. Například pět nejfrekventovanějších instrumentálů plurálu se sufixem *-ama* v korpusu SYN jsou podle frekvenční distribuce na výsledek dotazu `[word=".*ama" & tag="N..P7.*"]` tvary *klukama*, *holkama*, *Obama*, *Panama*, *Narama* (sportovní klub). Spolehlivé vyhodnocení frekvenčního vývoje sufixu *-ama* by tedy kromě zadání dotazu pouze na slovní tvary a vyřazení všech původně duálových tvarů vyžadovalo i odfiltrování dalších podobných proprií.

Také u přechodníku přítomného nastává obdobná situace způsobená velkým množstvím lexikalizovaných tvarů (*počínaje*, *konče*, *nemluvě* apod.), které by se ze synchronního hlediska za přechodníky považovat neměly. U přechodníku minulého (je tak značkován i přechodník přítomný tvořený od dokonavých sloves) je naproti tomu problémem jeho vzácnost, kvůli níž zůstává mezi jeho tvary relativně velké množství chyb a překlepů ve zdrojových textech: „dovolená *umoře*“, „nápad mi *připadna* zajímavý“, „*rozděle* ní vlivných postů“ atd.

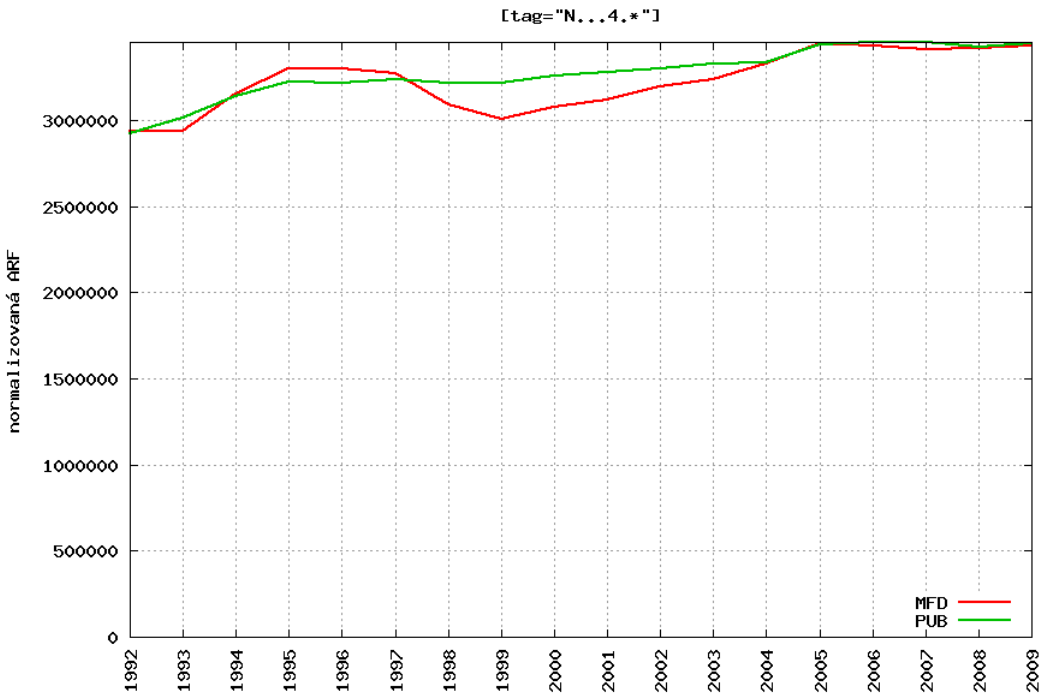
V některých případech jsou však spolehlivé a průkazné už výsledky jednoduchých dotazů na slovní tvary, což se týká například sufixu *-uju* (obr. 6.6.3). Přestože frekvence konkurenčního *-uji* neklesá (obr. 6.6.4), jde o zřejmý trend. Podobná situace je i u dvojice *-ují*/*-ujou*, ovšem s tím rozdílem, že varianta *-ujou* (obr. 6.6.5) je vzhle-

dem k frekvenci *-ují* (obr. 6.6.6) marginální a že její nárůst není tak pravidelný, což je způsobeno pravděpodobně její nízkou frekvencí. Dalším nehomonymním sufixem je adjektivní sufix *-ýho*, jehož frekvenční nárůst lze charakterizovat také jako nepravidelný a pohybující se v oblasti s velice s nízkou frekvencí (obr. 6.6.7). Ačkoli konkurenční spisovný sufix *-ého* vykazuje mírný, ale stabilní pokles (obr. 6.6.8), nelze ho ani v tomto případě interpretovat jako ústup spisovné varianty, a to vzhledem ke stále zanedbatelné frekvenci sufixu *-ýho*.

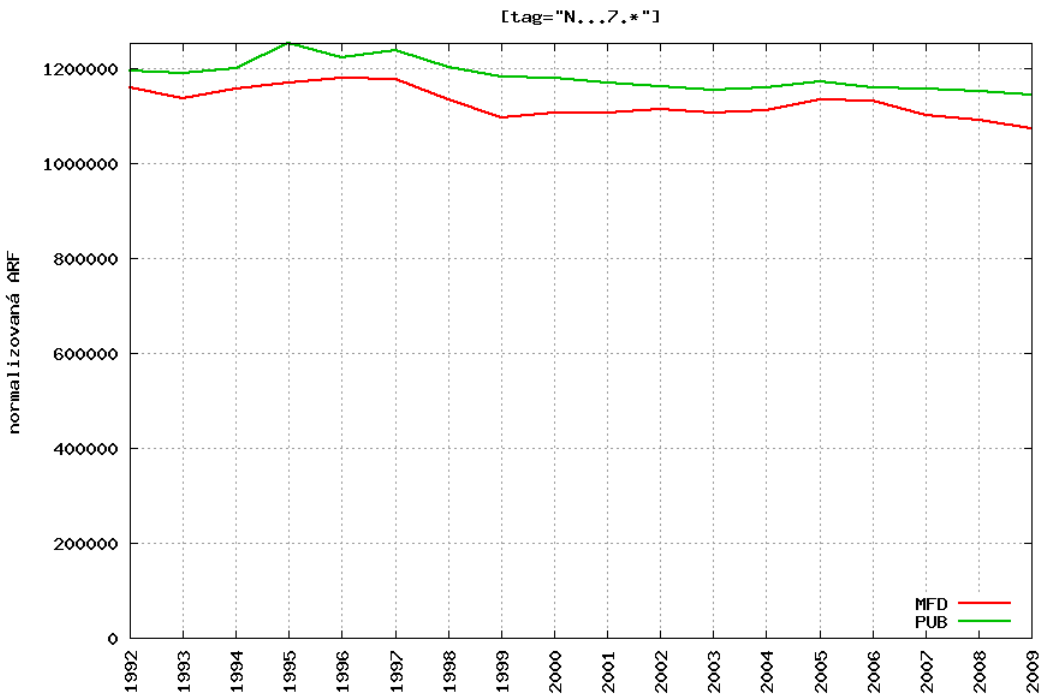
Výrazný nárůst vykazuje frekvenční průběh tvarů infinitivu končícího na *-ct* (obr. 6.6.9); ten není vyvážen poklesem tvarů končících na *-ci* (obr. 6.6.10), což je dáno celkovým frekvenčním nárůstem příslušných sloves (viz podkapitola 6.5). Celková frekvence infinitivů na *-ct* však za sledovaných 18 let vzrostla čtyřnásobně a dosáhla v roce 2009 poloviny frekvence infinitivů na *-ci*, celkový trend je tedy v tomto případě zřejmý a průkazný.

Postupný odklon od formálního vyjadřování lze ukázat také na klesající frekvenci opisného pasiva (obr. 6.6.11) a jmenných tvarů adjektiva (obr. 6.6.12; z dotazu byla vyloučena lemmata *rád* a *práv*, viz podkapitola 6.3). Projevuje se také klesajícím množstvím abstraktních feminin se sufixem *-ost* (obr. 6.6.13) a *-ace* (obr. 6.6.14).

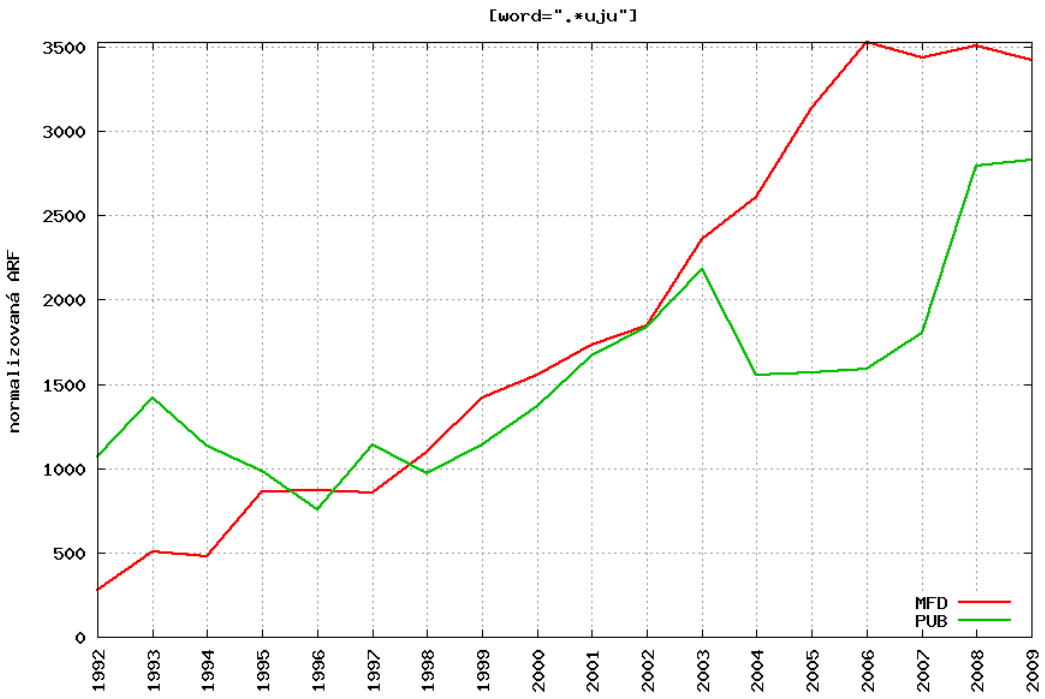
Z uváděných příkladů je zřejmé, že i přes zmíněné nedostatky zdrojových dat lze některé vývojové tendence zaznamenat poměrně přesvědčivě. Je však potřeba zdůraznit, že výpovědní hodnotu všech podrobnějších analýz by zvýšily některé navrhované změny v anotaci a složení korpusů, především doplnění dat v publicistice a zavedení jejího podrobnějšího členění jako korektivu případných nezáměrných posunů.



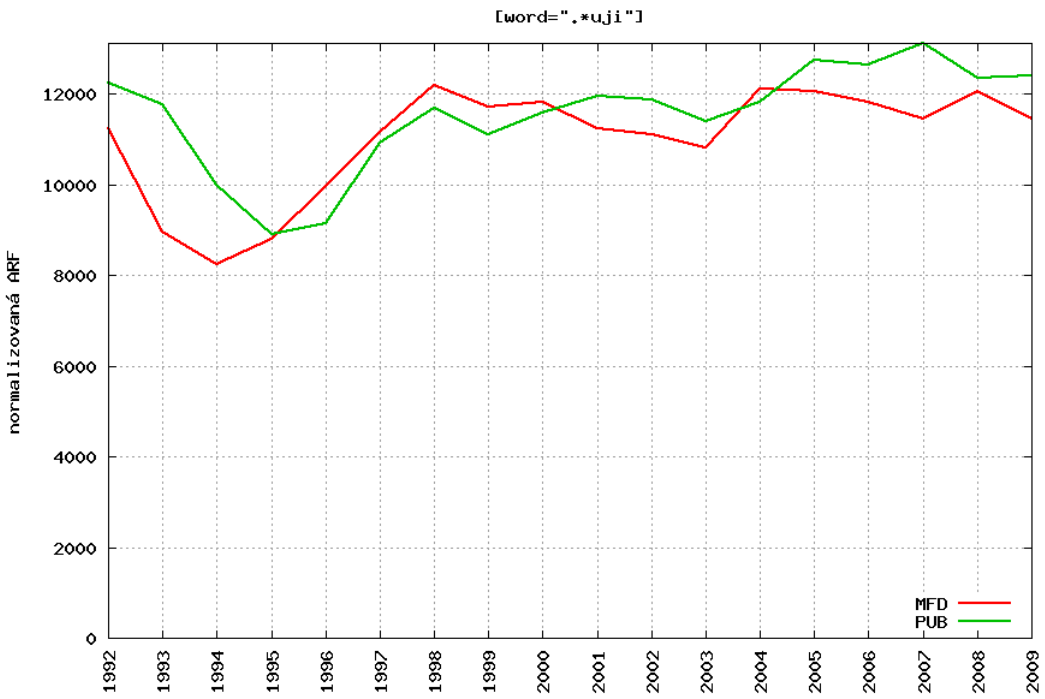
Obrázek 6.6.1: Průběh normalizované ARF akuzativu u substantiv.



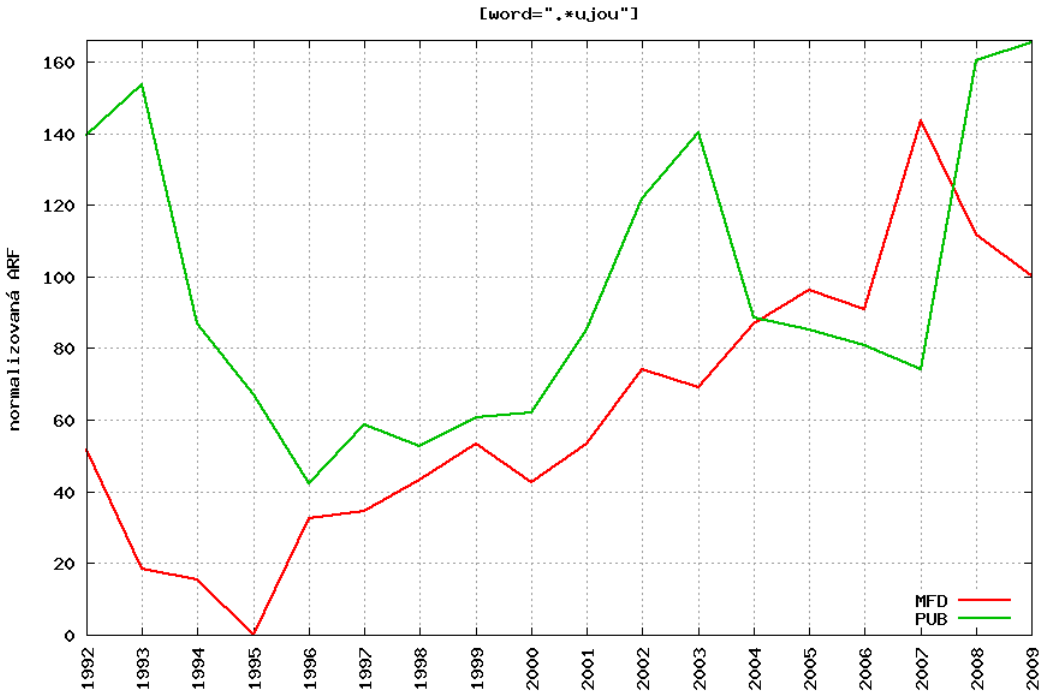
Obrázek 6.6.2: Průběh normalizované ARF instrumentálu u substantiv.



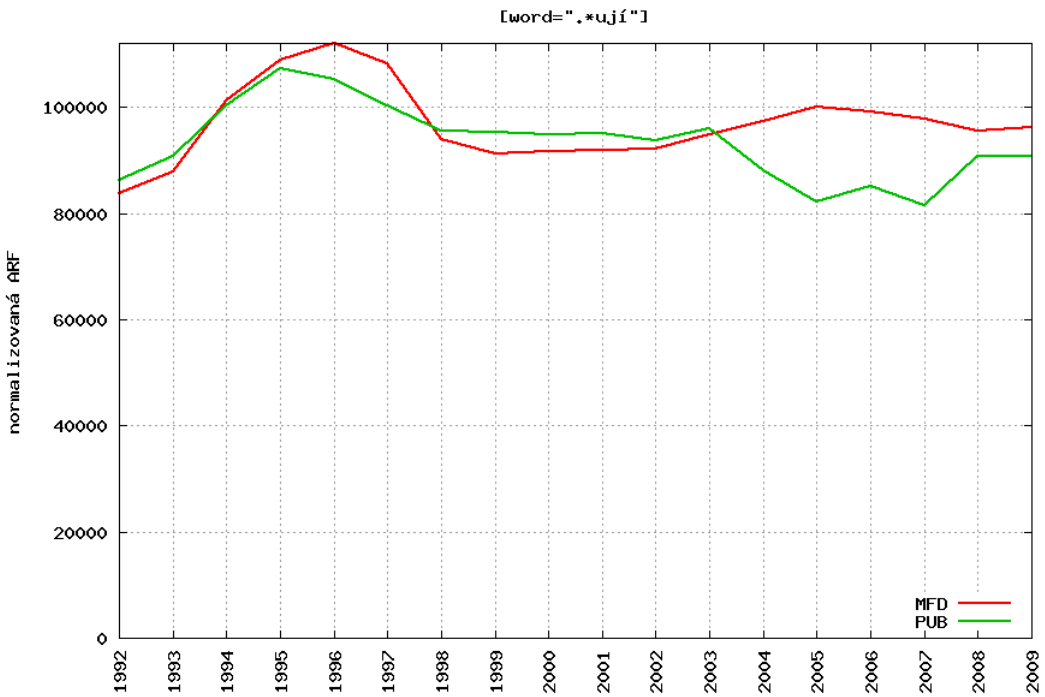
Obrázek 6.6.3: Průběh normalizované ARF tvarů se sufixem *-uju*.



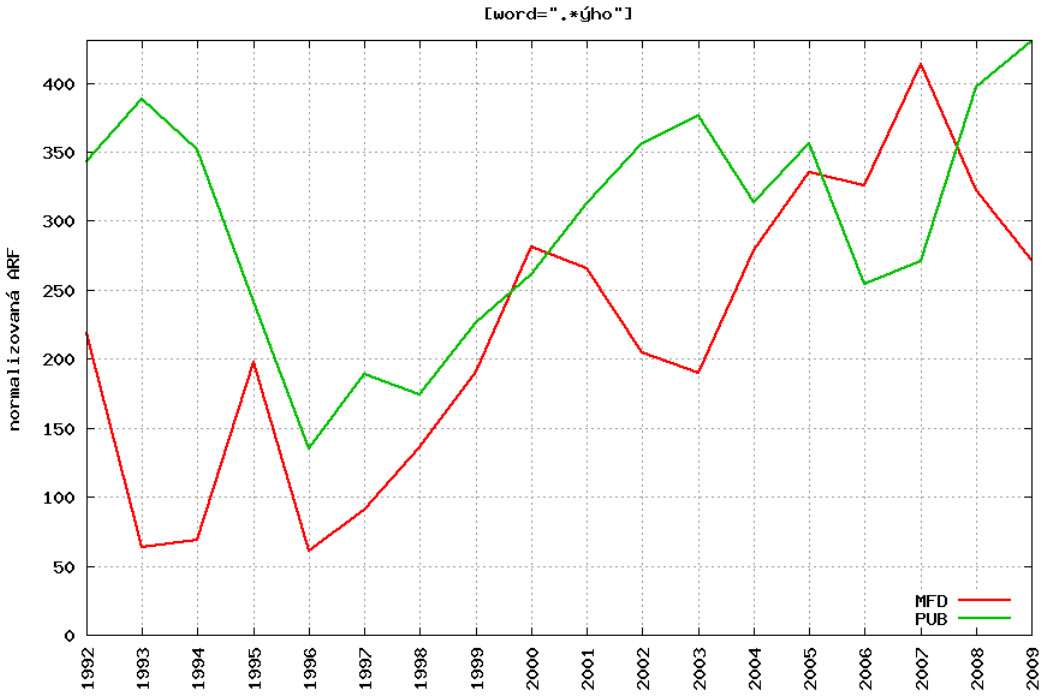
Obrázek 6.6.4: Průběh normalizované ARF tvarů se sufixem *-uji*.



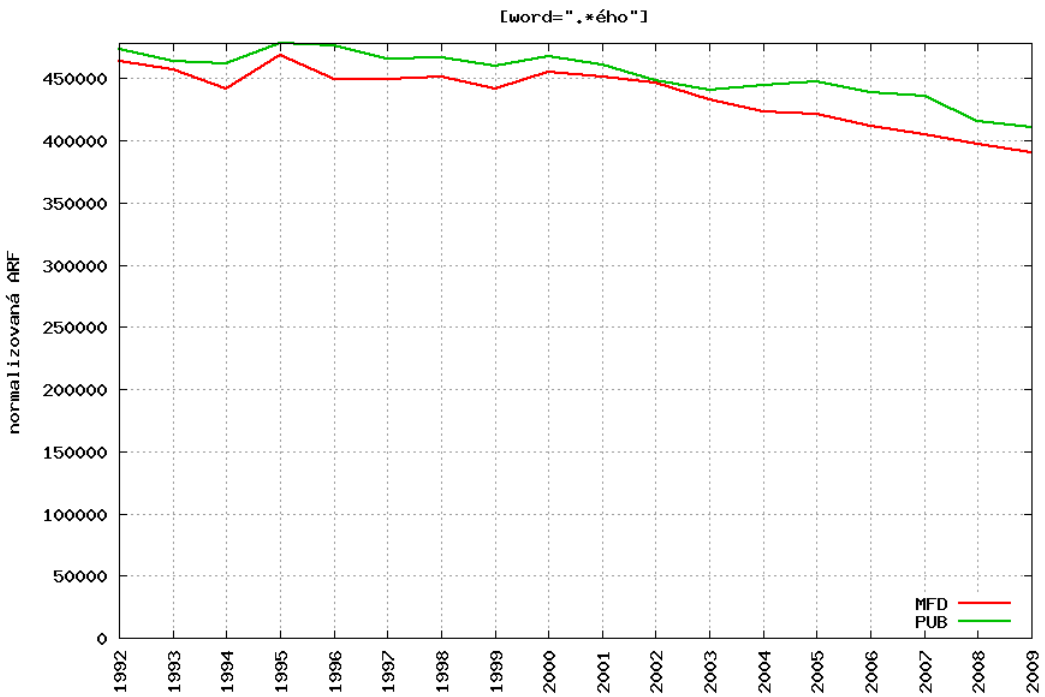
Obrázek 6.6.5: Průběh normalizované ARF tvarů se sufixem *-ujou*.



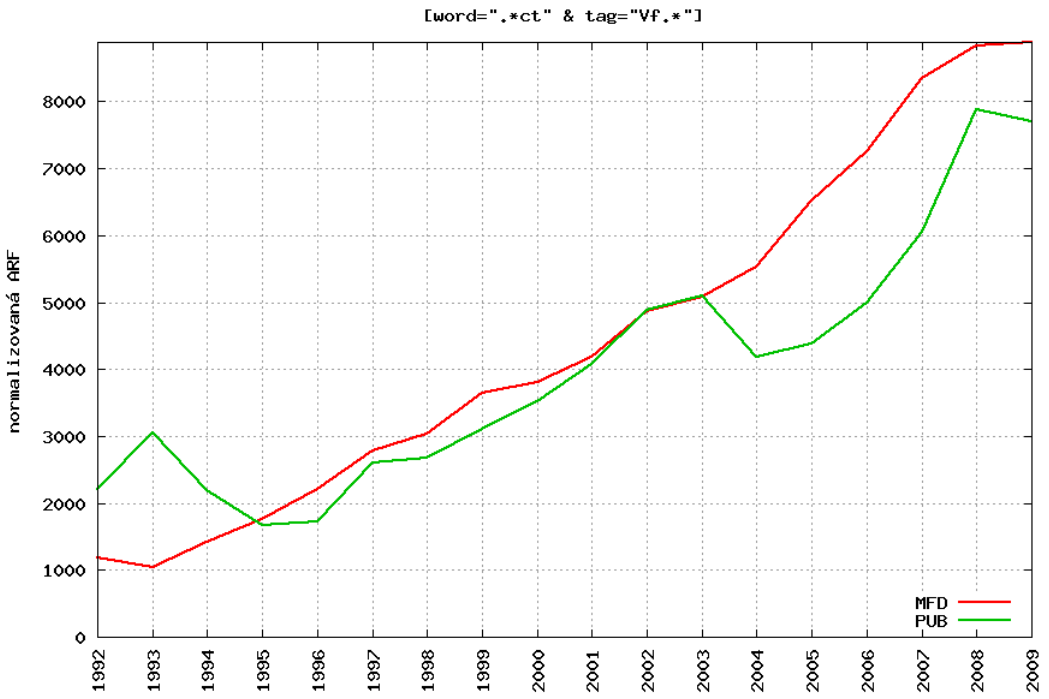
Obrázek 6.6.6: Průběh normalizované ARF tvarů se sufixem *-ují*.



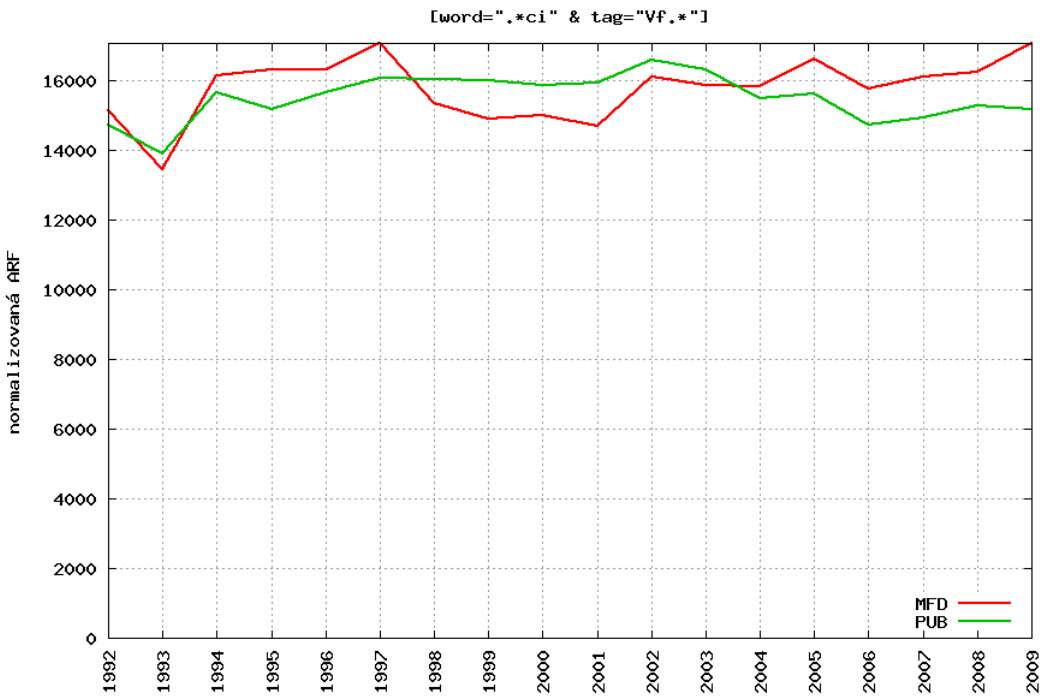
Obrázek 6.6.7: Průběh normalizované ARF tvarů se sufixem *-ýho*.



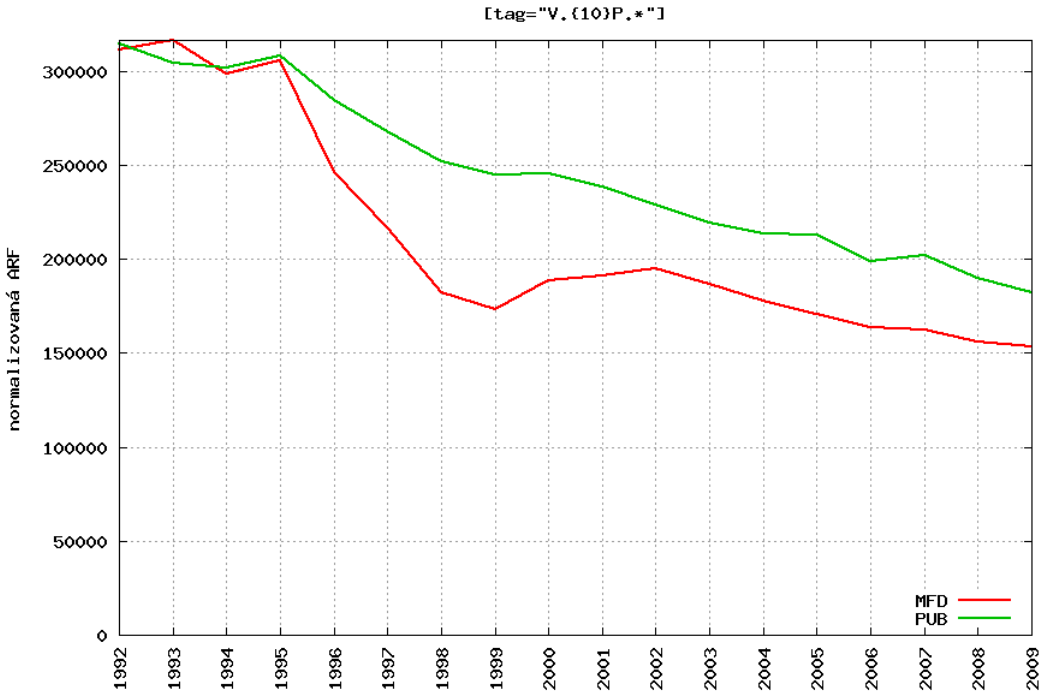
Obrázek 6.6.8: Průběh normalizované ARF tvarů se sufixem *-ého*.



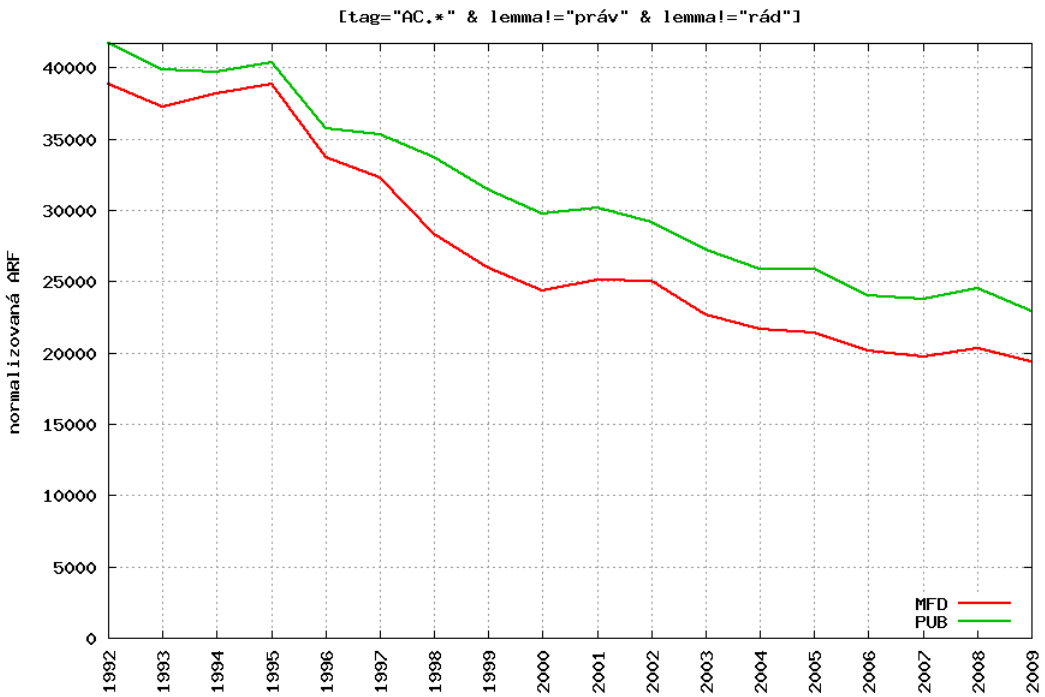
Obrázek 6.6.9: Průběh normalizované ARF tvarů infinitivu končícího na *-ct*.



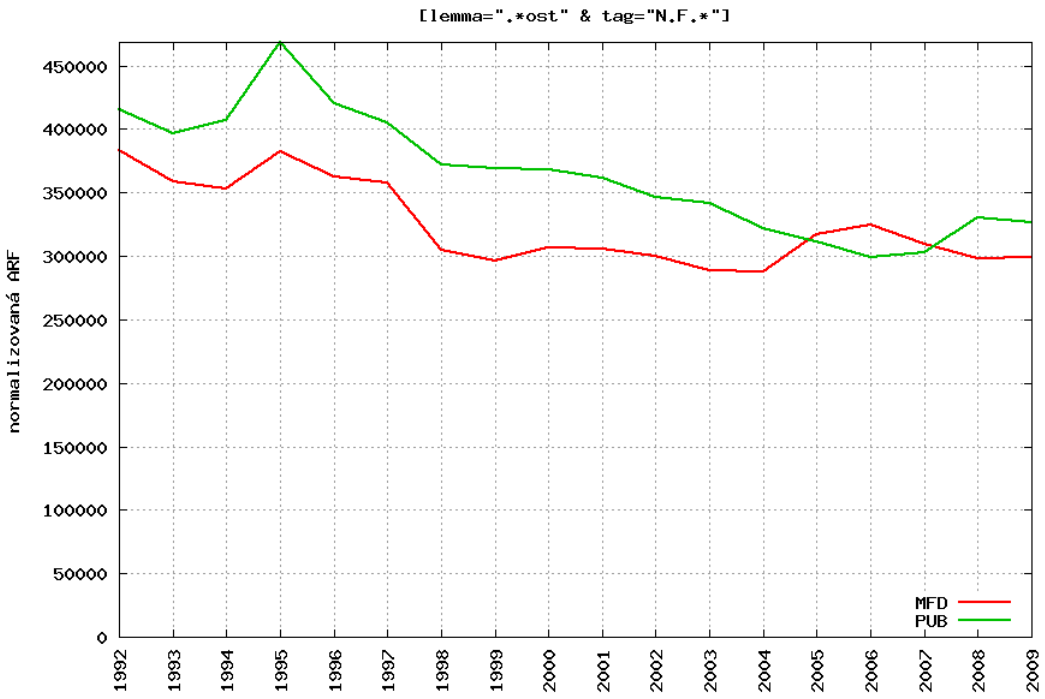
Obrázek 6.6.10: Průběh normalizované ARF tvarů infinitivu končícího na *-ci*.



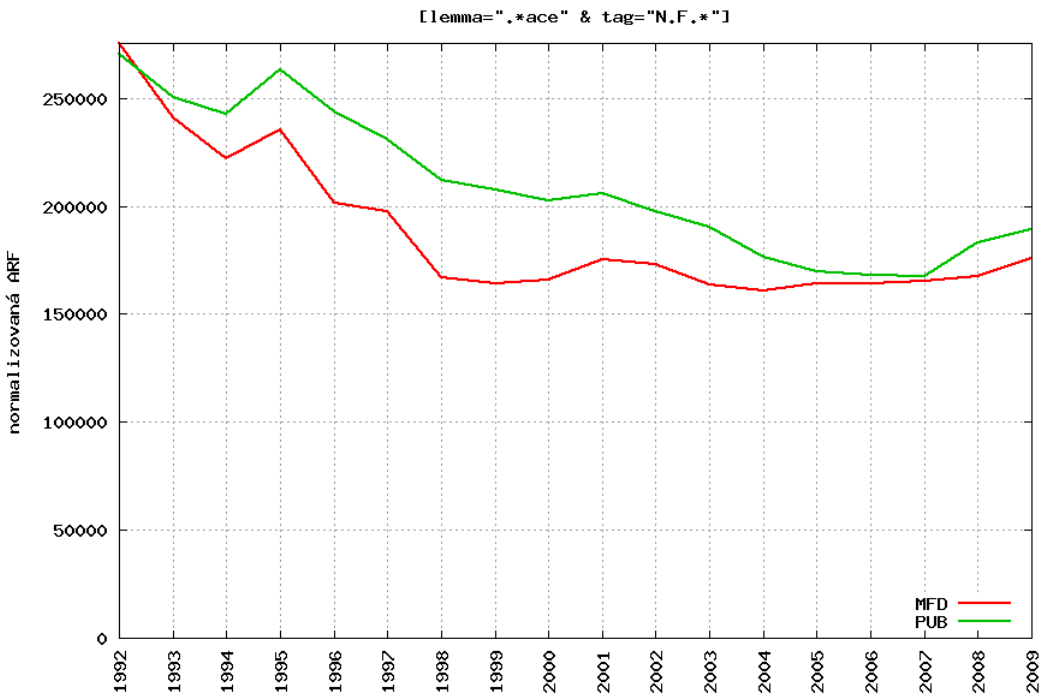
Obrázek 6.6.11: Průběh normalizované ARF tvarů slovesného pasiva.



Obrázek 6.6.12: Průběh normalizované ARF jmenných tvarů adjektiva.



Obrázek 6.6.13: Průběh normalizované ARF feminin se sufixem *-ost*.



Obrázek 6.6.14: Průběh normalizované ARF feminin se sufixem *-ace*.

7 Závěr

V práci byla rozpracována a vyhodnocena metoda pro diachronní srovnání synchronních korpusů zachycujících blízké stavy jazyka. Metodologicky jde o corpus-driven přístup založený na srovnávání frekvencí lemmat a lexikálních kombinací, jehož cílem je především zhodnotit možnosti a meze detekce vývojových tendencí v jazyce na materiálu synchronních psaných korpusů řady SYN. Návrhu použité metody předcházel popis vzniku, anotace a složení těchto korpusů a také přehled diachronních srovnání provedených na mnoha jiných korpusech. Metoda byla aplikována v několika variantách na různě definované subkorpusy korpusu SYN a podrobně vyhodnocena s výsledky poukazujícími nejenom na nedokonalé složení a způsob anotace zdrojových korpusů, ale především na zjištěné tendence jazykového vývoje.

Provedené srovnání ztěžují především blízké stavy jazyka, a to zvláště kvůli výraznému vlivu složení jednotlivých korpusů. Tento vliv je příliš velký zejména v případech, kdy je frekvence pozorovaných výrazů nízká nebo kdy jsou její změny pozvolné. V synchronním „šumu“ vzniklém nejenom rozdílným složením dat, ale také provázaností se změnami ve společnosti a společenskými tématy dané doby, tak může málo výrazná vývojová tendence snadno zaniknout. Bylo proto potřeba najít takový způsob diachronního srovnání, který minimalizuje vliv (nezáměrných) rozdílů ve složení srovnávaných korpusů. Metoda je proto založena na normalizované ARF používané namísto prosté frekvence, statisticky zjištěná významnost frekvenčních rozdílů jednotlivých výrazů je navíc zpětně ověřována na korpusech a interpretace výsledků korigována znalostí jejich přesného složení.

Jsme si vědomi zřejmého omezení použité metody vyplývající ze zvoleného způsobu srovnávání, které se provádí pouze na lexikální úrovni a na úrovni lexikálních kombinací. Jde o způsob vhodný především pro detekci neologismů nebo obecně výrazů, které vykazují výrazné frekvenční změny v úzu, což se týká také některých kolokací. Pro lexikon jsou však běžnější změny týkající se významu, jeho posunů a polysémie, které lze podobným způsobem zjistit pouze nepřímo, což platí o to více mimo lexikon, například pro morfologii nebo syntax.

Na druhou stranu lze očekávat, že při srovnávání velice blízkých stavů jazyka bude většina pozorovaných změn omezena právě na lexikon a jeho kombinatoriku, změny v jiných oblastech nebudou příliš znatelné. Posuny v jazyce mimo lexikální úroveň však přesto vysledovat lze, a jak ukazují některé části diskuse v kapitole 6, může na

řadu z nich poukázat studium výsledků metody aplikované pouze na lexikální úrovni. Práce proto naznačuje možné směry dalšího studia jazykových změn v publicistice, jde zejména o podrobné rozpracování řady corpus-based studií naznačených v podkapitole 6.6.

Poznamenejme také, že veškerá očekávání týkající se výsledků práce musejí být založena na tom, co se dá zjistit z dat, která jsou k dispozici. Znamená to například, že nelze očekávat nalezení vývojových tendencí typických pouze pro mluvený jazyk, ačkoli právě ten je nositelem většiny změn. Práce je však založena výhradně na jazyce psaném, nepočítáme-li rozhovory a přímé řeči v beletrii a publicistice, v nichž je ale mluvený jazyk zastoupen jen zprostředkovaně a v malém množství.

Prozatím není zřejmé, do jaké míry jsou měnící se stylistické konvence v publicistice ovlivňovány jazykem mluveným, ani jak (a zda vůbec) tyto změny v konečném důsledku vedou ke změnám v jazyce. Protože neumíme spolehlivě odlišit diachronní posuny od přirozeně existující synchronní variability, může práce sloužit nejenom jako korektiv našich očekávání od sledování frekvenčních průběhů lemmat a jejich kombinací v relativně krátkém časovém období, ale lze ji považovat také za specifický příspěvek k poznání jazykové variability.

Současné složení reprezentativních korpusů řady SYN není příliš vhodné pro diachronní srovnání blízkých stavů jazyka. Vývoj psaného jazyka jako celku po roce 1990 je tedy zatím korpusově obtížně uchopitelný, musíme se proto omezit na publicistiku, čímž se ale nutně snižuje vypovídací hodnota výsledků víceméně pouze na mediální diskurs, ačkoli výsledky založené na homogennějších datech jsou mnohem spolehlivější.

Veškeré závěry týkající se jazyka jsou proto založeny pouze na publicistice, která je však z psaného jazyka nejvíce otevřená změnám; ostatní typy textu jsou buď konzervativnější (odborná literatura), nebo se v nich projevuje větší setrvačnost (beletrie). Tento jev přímo studují Hundt a Mair (1999), podobná tvrzení se ale objevují i v dalších pracích (Křen a Hlaváčová, 2008; Millar, 2009).

Na základě publicistických subkorpusů byly popsány změny v jazyce publicistiky, jejichž charakteristiku je možné shrnout do následujících bodů:

1. publicistika je proměnlivým obrazem doby, reality a společenských témat, její jazyk věrně odráží jak periodicitu či naopak nepravidelnost konkrétních událostí, tak i nástup nových technologií a proměny životního stylu;

2. publicistika se odklání od původní politické a ekonomické orientace směrem k tématům týkajícím se praktického života a využívání volného času;

3. zvyšující se neformálnost jazyka publicistiky způsobuje posuny ve frekvencích některých slovních druhů, frekvenční nárůst řady lemmat z jádra slovní zásoby, vzrůstající podíl významově oslabených sloves v kategoriálním užití, obměnu některých šablonovitých spojení atd.

Některé z těchto charakteristik odpovídají výsledkům zjištěným na angličtině v rámci studií MS-CADS zmíněných již v oddílu 3.3.3. Zvětšující se rozsah jednotlivých čísel MFD spojený především s rostoucím počtem víkendových a jiných zájmových příloh, který jsme konstatovali v podkapitole 4.5, koresponduje s podobnými zjištěními týkajícími se britského tisku. To by však umožňovalo vysvětlit vzrůstající neformálnost publicistiky pouhým přibýváním velkého množství textů jiného žánru. Jde jistě o jednu z příčin, Duguid (2010, str. 134) však konstatuje, že k těmto posunům dochází i u titulů, jejichž objem se nezvětšil, a že tedy pravděpodobně jde o kombinaci více faktorů. Toto tvrzení bohužel nelze ověřit na MFD, protože není k dispozici její vnitřní členění na jednotlivé tematicky zaměřené části.

Duguid (2010, str. 113–114) uvádí kromě neformálnosti také další charakteristiky, například zvyšující se pozitivní evaluativnost. Jako jednu z příčin zmiňuje rozšiřující se přebírání již hotových textů, hlavně z public relations velkých korporací; tomu z našich zjištění odpovídá rostoucí obliba citování tiskových mluvčích. Duguid (2010, str. 117) dále zdůrazňuje roli přímé řeči, některé jazykové prostředky mohou být omezeny pouze na ni, nemusel je tedy použít přímo novinář. Přesto je však pochopitelně významný sám fakt, že se v publicistice tyto prostředky, byť zprostředkovaně, vůbec objevují.

Zmiňovaná vzrůstající neformálnost publicistiky je však zřejmá nejenom na lexikální a morfologické úrovni, posuny v zastoupení jednotlivých slovních druhů naznačují posun od odborného stylu směrem k beletrii, což v zásadě odpovídá jejímu přibližování jazyku mluvenému, přinejmenším v některých rysech (nasvědčuje mu i zjištěný frekvenční nárůst většiny frekventovaných lemmat). To se shoduje s tzv. „colloquialisation“ popisovanou jako „tendency for the written language gradually to acquire norms and characteristics associated with the spoken conversational language“ (Leech, 2004, str. 72); na tento jev poukazuje také řada dalších studií (mj. Baker, 2009a; Duguid, 2010; Hundt a Mair, 1999; Leech, 2003; Mair, 1997; Mair et al., 2002; Millar, 2009).

Protože neformální způsob vyjadřování nevzniká nově, ale stává se „pouze“ přijatelným v dalších situacích a kontextech, je tato neformálnost jevem spíše socio-kulturním než jazykovým. Přesto je mezi nimi zřejmý vztah: „ ... in due course, it will no doubt have consequences for the linguistic system, because the new stylistic climate will speed up the demise of many lexical and grammatical archaisms and prevent the establishment of new lexical and grammatical markers of more formal or literary diction.“ (Mair et al., 2002, str. 256)

Otevřenou otázkou zůstává, zda uvedené závěry založené na publicistických subkorpusech nelze alespoň v omezené míře vztáhnout na psaný jazyk nebo dokonce na češtinu jako celek. Vzhledem k nejednoznačnému vztahu mezi korpusem a jazykovou realitou je zřejmé, že zobecňování na korpuse založených závěrů na jazyk nebo jeho konkrétní varietu by mělo být opatrné i v případě, že jde o varietu, kterou by měl daný korpus přímo reprezentovat. Domníváme se proto, že nás k takovému zobecnění

nic neopravňuje, ačkoli je pravděpodobné, že zvyšující se neformálnost publicistiky je způsobena celospolečenským klimatem, jehož působení se neomezuje pouze na ni.

Tématem vhodným pro budoucí práci je adaptace použité corpus-driven metody a její převedení z lexikální úrovně na gramatickou, konkrétně na identifikaci kombinací morfologických kategorií s nejvýrazněji nebo nejpravidelněji se měnícím frekvenčním průběhem v čase. Na obecnější rovině by podrobnější studii zasloužily posuny v postavení publicistiky v rámci ostatních hlavních typů textu založené na komplexnější množině jazykových jevů vycházející například z Biberovy vícerozměrné analýzy (viz podkapitola 2.3).

Praktickým výstupem práce se může stát mechanismus tvorby grafů použitý v kapitole 6, který je po dalších úpravách a doplnění zdrojových dat možné zveřejnit jako webovou službu, a umožnit tak tento způsob práce s publicistickými subkorpusy i veřejnosti. Jeho výhodou by byla možnost zadávat velice obecné dotazy na data ze sice omezeného časového období, ale s vysokou mírou spolehlivosti poskytovaných výsledků.

Kromě teoretického přínosu ke korpusovému popisu jazykových změn bychom chtěli závěrem zdůraznit potřebnost práce pro vyhodnocení složení korpusů řady SYN, zvláště reprezentativních korpusů SYN2000, SYN2005 a SYN2010. Pokud je nám známo, tyto korpusy dosud podobným způsobem srovnány nebyly, a to ani interně v ÚČNK. Mezi výsledky práce je řada praktických doporučení ke změnám v konceptu reprezentativnosti, kategorizaci textů a složení korpusových dat, která byla shrnuta v podkapitole 6.6 a která tvoří cennou zpětnou vazbu pro budování dalších korpusů této řady. Tato doporučení nevycházejí pouze z potřeb diachronní srovnatelnosti korpusů řady SYN, řadu z nich by podle našeho názoru uvítal také širší okruh uživatelů ČNK.

Použité korpusy

Český národní korpus – SYN. Verze z 20. prosince 2010. Ústav Českého národního korpusu FF UK Praha. URL: `http://www.korpus.cz`.

Český národní korpus – SYN2000. Ústav Českého národního korpusu FF UK Praha. URL: `http://www.korpus.cz`.

Český národní korpus – SYN2005. Ústav Českého národního korpusu FF UK Praha. URL: `http://www.korpus.cz`.

Český národní korpus – SYN2006PUB. Ústav Českého národního korpusu FF UK Praha. URL: `http://www.korpus.cz`.

Český národní korpus – SYN2009PUB. Ústav Českého národního korpusu FF UK Praha. URL: `http://www.korpus.cz`.

Český národní korpus – SYN2010. Ústav Českého národního korpusu FF UK Praha. URL: `http://www.korpus.cz`.

Literatura

Alderson, J. C. (2007). "Judging the frequency of English words". In: *Applied Linguistics* 28.3, pp. 383–409.

Altintas, K., F. Can a J. M. Patton (2007). "Language change quantification using time-separated parallel translations". In: *Literary and Linguistic Computing* 22.4, pp. 375–393.

Asmussen, J. (2006). "Towards a methodology for corpus-based studies of linguistic change: Contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 General Corpora of Danish". In: *Corpus Linguistics Around the World*. Ed. D. Archer, P. Rayson a A. Wilson. Amsterdam: Rodopi, pp. 33–48.

Atkins, S., J. Clear a N. Ostler (1992). "Corpus design criteria". In: *Literary and Linguistic Computing* 7.1, pp. 1–16.

Baker, P. (2009a). "The BE06 Corpus of British English and recent language change". In: *International Journal of Corpus Linguistics* 14.3, pp. 312–337.

Baker, P. (2009b). "The British English '06 Corpus – using the LOB model to build a contemporary corpus from the internet". In: *Proceedings of the Corpus Linguistics Conference*. Ed. M. Mahlberg, V. González-Díaz a C. Smith. Liverpool.

Banerjee, S. a T. Pedersen (2003). "The Design, Implementation and Use of the Ngram Statistics Package". In: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, pp. 370–381.

Baroni, M., A. Kilgarriff, J. Pomikálek a P. Rychlý (2006). "WebBootCaT: a web tool for instant corpora". In: *Proceedings of the 12th EURALEX International Congress*. Ed. E. Corino, C. Marello a C. Onesti. Torino, pp. 123–131.

Baroni, M., S. Bernardini, A. Ferraresi a E. Zanchetta (2009). "The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora". In: *Journal of Language Resources and Evaluation* 43.3, pp. 209–226.

Bartoň, T., V. Cvrček, F. Čermák, T. Jelínek a V. Petkevič (2009). *Statistiky češtiny*. Praha: NLN.

Belica, C. (1996). "Analysis of Temporal Changes in Corpora". In: *International Journal of Corpus Linguistics* 1.1, pp. 61–73.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). "Representativeness in Corpus Design". In: *Literary and Linguistic Computing* 8.4, pp. 243–257.

Biber, D. (1995). *Dimensions of register variation. A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D., S. Conrad a R. Reppen (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D. a E. Finegan (2001). "Diachronic relations among speech-based and written registers in English". In: *Variation in English: Multi-Dimensional Studies*. Ed. S. Conrad a D. Biber. London: Longman.

Biber, D., E. Finegan a D. Atkinson (1994). "ARCHER and its challenges: Compiling and exploring 'A Representative Corpus of Historical English Registers'". In: *Creating and using English language corpora*. Ed. U. Fries, G. Tottie a P. Schneider. Amsterdam: Rodopi, pp. 1–14.

Biber, D., S. Johansson, G. Leech, S. Conrad a E. Finegan (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Čermák, F. (1997). "Czech National Corpus: A Case in Many Contexts". In: *International Journal of Corpus Linguistics* 2.2, pp. 181–197.

Čermák, F. (2001a). "Language corpora: the Czech case". In: *Text, Speech, Dialogue. 4th International Conference Proceedings*. Ed. V. Matoušek, P. Mautner, R. Mouček a K. Taušer. Berlin: Springer, pp. 21–30.

Čermák, F. (2001b). "Syntagmatika slovníku: typy lexikálních kombinací". In: *Čeština – univerzália a specifika 3*. Ed. Z. Hladká a P. Karlík. Brno: MU, pp. 223–232.

Čermák, F. (v tisku). "Centrum a periférie ještě jednou". In: *Človek a jeho jazyk 3. Inšpirácie profesora Jána Horeckého*. Ed. M. Šimková. Bratislava: Veda.

Čermák, F., V. Cvrček et al. (2009). *Slovník Bohumila Hrabala*. Praha: NLN.

Čermák, F., J. Králík a K. Kučera (1997). "Recepce současné češtiny a reprezentativnost korpusu". In: *Slovo a slovesnost* 58.2, pp. 117–124.

Čermák, F., M. Křen et al. (2004). *Frekvenční slovník češtiny*. Praha: NLN.

Čermák, F., M. Křen et al. (2011). *A Frequency Dictionary of Czech. Core Vocabulary for Learners*. London: Routledge.

Čermák, F. et al. (2007). *Slovník Karla Čapka*. Praha: NLN.

Cvrček, V. a P. Vondřička (2011). "Výzkum variability v korpusech češtiny". In: *Korpusová lingvistika Praha 2011. 2 Výzkum a výstavba korpusů*. Ed. F. Čermák. Praha: NLN, pp. 184–195.

Davies, M. (2009). "The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights". In: *International Journal of Corpus Linguistics* 14.2, pp. 159–190.

Davies, M. (2010). "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English". In: *Literary and Linguistic Computing* 25.4, pp. 447–464.

Davies, M. (2011). "Examining Recent Changes in English: Some Methodological Issues". In: *Handbook on the History of English: Rethinking Approaches to the History of English*. Ed. T. Nevalainen a E. C. Traugott. Oxford: Oxford University Press.

Duguid, A. (2010). "Newspaper discourse informalisation: a diachronic comparison from keywords". In: *Corpora* 5.2, pp. 109–138.

Dunning, T. (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". In: *Computational Linguistics* 19.1, pp. 61–74.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Stuttgart: University of Stuttgart.

Fletcher, W. H. (2004). "Making the Web more useful as a source for linguistic corpora". In: *Corpus Linguistics in North America 2002*. Ed. U. Connor a T. Upton. Amsterdam: Rodopi, pp. 191–205.

Gries, S. T. (2008). "Dispersions and adjusted frequencies in corpora". In: *International Journal of Corpus Linguistics* 13.4, pp. 403–437.

Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha: Karolinum.

Hanks, P. (2000). "Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance". In: *Proceedings of the 9th EURALEX International Congress*. Ed. U. Heid, S. Evert, E. Lehmann a C. Rohrer. Stuttgart, pp. 3–13.

Hilpert, M. a S. T. Gries (2009). "Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition". In: *Literary and Linguistic Computing* 24.4, pp. 385–401.

Hofland, K. a S. Johansson (1982). *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.

Hundt, M. a G. Leech (2011). "Small is beautiful – on the value of standard reference corpora for observing recent grammatical change". In: *Handbook on the History of English: Rethinking and Extending Approaches and Methods*. Ed. T. Nevalainen a E. C. Traugott. Oxford: Oxford University Press.

Hundt, M. a C. Mair (1999). "'Agile' and 'Uptight' Genres: The Corpus-based Approach to Language Change in Progress". In: *International Journal of Corpus Linguistics* 4.2, pp. 221–242.

Jelínek, T. (2008). "Nové značkování v Českém národním korpusu". In: *Naše řeč* 91.1, pp. 13–20.

Johansson, S. a K. Hofland (1989). *Frequency Analysis of English Vocabulary and Grammar*. Oxford: Clarendon.

Kilgarriff, A. (2001). "Comparing Corpora". In: *International Journal of Corpus Linguistics* 6.1, pp. 97–133.

Kocek, J., M. Kopřivová a K. Kučera (2000). *Český národní korpus – Úvod a příručka uživatele*. Praha: FF UK.

Králík, J. (2001). "Vyvážení zdrojů Synchronního korpusu češtiny SYN2000". In: *Slovo a slovesnost* 62.1, pp. 38–53.

Králík, J. (2004). "Aktualizace rozvržení zdrojů Českého národního korpusu s ohledem na revizi vyváženosti jeho struktury". In: *Slovo a slovesnost* 65.2, pp. 133–142.

Králík, J. (2006). "Zamyšlení nad velkými výběry". In: *Korpusová lingvistika: Stav a modelové přístupy*. Ed. F. Čermák a R. Blatná. Praha: NLN, pp. 205–209.

Králík, J. a M. Šulc (2005). "The Representativeness of Czech Corpora". In: *International Journal of Corpus Linguistics* 10.3, pp. 357–366.

Křen, M. (2006a). "Frequency Dictionary of Czech: A Detailed Processing Description". In: *Insight into the Slovak and Czech Corpus Linguistics*. Ed. M. Šimková. Bratislava: Veda, pp. 16–25.

Křen, M. (2006b). "Kolokační míry a čeština: srovnání na datech Českého národního korpusu". In: *Kolokace*. Ed. F. Čermák a M. Šulc. Praha: NLN, pp. 223–248.

Křen, M. (2006c). "SYN2000 vs. SYN2005: Comparing the Large Synchronic Corpora of Czech". In: *Proceedings of the International Conference "Corpus linguistics – 2006"*. St. Petersburg: St. Petersburg University Press, pp. 182–189.

Křen, M. (2007). "Variation of Czech Lexicon as Reflected by Corpora Comparison". In: *Computer Treatment of Slavic and East European Languages*. Bratislava: Tribun, pp. 109–120.

Křen, M. (2008). "Compilation of the Dictionary of Karel Čapek". In: *Corpus Linguistics, Computer Tools, and Applications – State of the Art*. Ed. B. Lewandowska-Tomaszczyk. Frankfurt am Main: Peter Lang, pp. 469–481.

Křen, M. (2009). "The SYN Concept: Towards One-Billion Corpus of Czech". In: *Proceedings of the Corpus Linguistics Conference*. Ed. M. Mahlberg, V. González-Díaz a C. Smith. Liverpool.

Křen, M. a J. Hlaváčová (2008). "Corpus as a Means for Study of Lexical Usage Changes". In: *Proceedings of the 13th EURALEX International Congress*. Ed. E. Bernal a J. DeCesaris. Barcelona, pp. 437–447.

Kučera, K. (2002). "The Czech National Corpus: Principles, Design, and Results". In: *Literary and Linguistic Computing* 17.2, pp. 245–257.

Kučera, K. (2005). "Diachronní kvantitativní pohled na vybrané případy konkurence v češtině a otázka smyslu budování diachronního korpusu". In: *Verba et historia*. Ed. P. Nejedlý a M. Vajdlová. Praha: ÚJČ AV ČR, pp. 191–196.

Kučera, K. (2011). "Diachronní složka Českého národního korpusu: historie, přítomnost, budoucnost". In: *Korpusová lingvistika Praha 2011. 2 Výzkum a výstavba korpusů*. Ed. F. Čermák. Praha: NLN, pp. 64–73.

Kytö, M. (1991). *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts*. Helsinki: Helsinki University Press.

Lee, D. (2001). "Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle". In: *Language Learning and Technology* 5.3, pp. 37–72.

Leech, G. (1991). "The state of the art in corpus linguistics". In: *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Ed. K. Aijmer a B. Altenberg. London: Longman, pp. 8–29.

Leech, G. (2003). "Modals on the move: The English modal auxiliaries 1961–1992". In: *Modality in Contemporary English*. Ed. R. Facchinetti, M. Krug a F. Palmer. Berlin: Mouton de Gruyter, 223–240.

Leech, G. (2004). "Recent grammatical change in English: Data, description, theory". In: *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*. Ed. K. Aijmer a B. Altenberg. Amsterdam: Rodopi, pp. 61–81.

Lew, R. (2009). "The Web as Corpus Versus Traditional Corpora: Their Relative Utility for Linguists and Language Learners". In: *Contemporary Corpus Linguistics*. Ed. P. Baker. London: Continuum, pp. 289–300.

Lieberman, E., J.-B. Michel, J. Jackson, T. Tang a M. A. Nowak (2007). "Quantifying the evolutionary dynamics of language". In: *Nature* 449, pp. 713–716.

Lüdeling, A., S. Evert a M. Baroni (2007). "Using web data for linguistic purposes". In: *Corpus Linguistics and the Web*. Ed. M. Hundt, N. Nesselhauf a C. Biewer. Amsterdam: Rodopi, pp. 7–24.

Mair, C. (1997). "Parallel corpora: A real-time approach to the study of language change in progress". In: *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)*. Ed. M. Ljung. Amsterdam: Rodopi, pp. 195–209.

Mair, C. (2011). "Change and diversification in present-day English: web-based perspectives". In: *Handbook on the History of English: Rethinking and Extending Approaches and Methods*. Ed. T. Nevalainen a E. C. Traugott. Oxford: Oxford University Press.

Mair, C., M. Hundt, G. Leech a N. Smith (2002). "Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora". In: *International Journal of Corpus Linguistics* 7.2, pp. 245–264.

Martínek, F. (2011). "Synchronní a diachronní pohled na jeden typ slovesné polysémie". In: *Korpusová lingvistika Praha 2011. 2 Výzkum a výstavba korpusů*. Ed. F. Čermák. Praha: NLN, pp. 262–272.

McGee, I. (2008). "Word Frequency Estimates Revisited – A Response to Alderson (2007)". In: *Applied Linguistics* 29.3, pp. 509–514.

Meurman-Solin, A. (2001). "Structured Text Corpora in the Study of Language Variation and Change". In: *Literary and Linguistic Computing* 16.1, pp. 5–27.

Michel, J.-B. et al. (2011). "Quantitative Analysis of Culture Using Millions of Digitized Books". In: *Science* 331.6014, pp. 176–182.

Millar, N. (2009). "Modal verbs in TIME: Frequency changes 1923–2006". In: *International Journal of Corpus Linguistics* 14.2, pp. 191–220.

Nelsen, R. B. (2001). "Kendall tau metric". In: *Encyclopaedia of Mathematics*. Ed. M. Hazewinkel. Berlin: Springer.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Oakes, M. (2009). "Corpus Linguistics and Language Variation". In: *Contemporary Corpus Linguistics*. Ed. P. Baker. London: Continuum, pp. 159–183.

Oakes, M. a M. Farrow (2007). "Use of the chi-squared test to examine vocabulary differences in English-language corpora representing seven different countries". In: *Literary and Linguistic Computing* 22.1, pp. 85–100.

Partington, A. S. (2010). "Modern Diachronic Corpus-Assisted Discourse Studies (MDCADS) on UK newspapers: an overview of the project". In: *Corpora* 5.2, pp. 83–108.

Pecina, P. (2009). *Lexical Association Measures: Collocation Extraction*. Vol. 4. Studies in Computational and Theoretical Linguistics. Praha: ÚFAL.

Petkevič, V. (2006). "Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary". In: *Insight into the Slovak and Czech Corpus Linguistics*. Ed. M. Šimková. Bratislava: Veda, pp. 26–44.

Rayson, P. a R. Garside (2000). "Comparing Corpora using Frequency Profiling". In: *Proceedings of the Workshop on Comparing Corpora, Annual Meeting of the ACL Archive*. Vol. 9, pp. 1–6.

Renouf, A., A. Kehoe a J. Banerjee (2005). "The WebCorp Search Engine: a holistic approach to Web text Search". In: *Proceedings of the Corpus Linguistics Conference*. Ed. P. Danielsson a M. Wagenmakers. Birmingham.

Rissanen, M. (1994). "The Helsinki Corpus of English Texts". In: *Corpora across the centuries. Proceedings of the First International Colloquium on English Diachronic Corpora*. Ed. M. Kytö, M. Rissanen a S. Wright. Amsterdam: Rodopi, pp. 73–80.

Rychlý, P. (2000). *Korpusové manažery a jejich efektivní implementace*. Brno: FI MU.

Rychlý, P. (2007). "Manatee/Bonito – A Modular Corpus Manager". In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno, pp. 65–70.

Savický, P. a J. Hlaváčová (2002). "Measures of Word Commonness". In: *Journal of Quantitative Linguistics* 9.3, pp. 215–231.

Sharoff, S. (2006). "Creating General-Purpose Corpora Using Automated Search Engine Queries". In: *Wacky! Working papers on the Web as Corpus*. Ed. M. Baroni a S. Bernardini. Bologna: GEDIT, pp. 63–98.

Sigley, R. (1997). "Text Categories and Where You Can Stick Them: A Crude Formality Index". In: *International Journal of Corpus Linguistics* 2.2, pp. 199–237.

Sinclair, J. (2005). "Corpus and Text – Basic Principles". In: *Developing Linguistic Corpora – a Guide to Good Practice*. Ed. M. Wynne. Oxford: Oxbow books. URL: `http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm`.

Spoustová, D., J. Hajič, J. Votrubec, P. Krbec a P. Květoň (2007). "The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech". In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*. Praha, pp. 67–74.

Šulc, M. (2001). "Tematická reprezentativnost korpusů". In: *Slovo a slovesnost* 62.1, pp. 53–61.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.

Waclawičová, M. a M. Křen (2008). "ORAL2008: New Balanced Corpus of Spoken Czech". In: *Proceedings of the International Conference "Corpus linguistics – 2008"*. St. Petersburg: St. Petersburg University Press, pp. 105–112.

Waclawičová, M., M. Křen a L. Válková (2009). "Balanced Corpus of Informal Spoken Czech: Compilation, Design and Findings". In: *Proceedings of the 10th Annual Conference of the International Speech Communication Association INTERSPEECH 2009*. Brighton: ISCA, pp. 1819–1822.