Osolsobě, Klára; Pala, Karel; Rychlý, Pavel

Frekvence vzorů českých substantiv : (na materiálu ČNK)

Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná. 1998, vol. 47, iss. A46, pp. [77]-94

ISBN 80-210-1796-1
ISSN 0231-7567

SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 46, 1998

KLÁRA OSOLSOBĚ, KAREL PALA, PAVEL RYCHLÝ

FREKVENCE VZORŮ ČESKÝCH SUBSTANTIV
(na materiálu ČNK)

1. Úvod

V tomto článku bychom rádi nabídli některé nové údaje o frekvencích substantivních vzorů a substantiv v současné češtině a porovnali je s údaji staršími, obsaženými ve Frekvenčním slovníku češtiny (FSČ,1961) a také zčásti v práci M. Těšitelové (1985). Naše výsledky vycházejí z již částečně vybudovaného Českého národního korpusu, konkrétně z jeho volně navazující podčásti — značkovaného subkorpusu s názvem DESAM (obsahuje 1 026 733 slovních tvarů), který byl vytvořen v rámcové spolupráci pracovišť sdružených v grantovém projektu K214, konkrétně pak na FF MU a FI MU v Brně (Komplexní grantový projekt *Čeština ve věku počítačů* se uskutečňuje díky finanční podpoře GA ČR a sdružuje 7 pracovišť: Ústav českého národního korpusu při FF UK (F. Čermák), Ústav formální a aplikované lingvistiky na MFF UK (E. Hajičová), Ústav formální a teoretické lingvistiky FF UK (P. Sgall, V. Petkevič), Katedra českého jazyka FF UK (K. Kučera), Ústav českého jazyka FF MU (K. Osolsobě), Katedra informačních technologií FI MU (K. Pala), Ústav českého jazyka AV ČR (J. Králík)). I když jde o předběžné výsledky, rozsah korpusu DESAM je podle našeho názoru dostatečný pro uvedení některých současných tendencí v české substantivní deklinaci a jejich porovnání se staršími zjištěními. Materiál korpusu DESAM poskytuje též podklady pro některé zajímavé postřehy ke změnám a tendencím ve slovní zásobě současné češtiny.

Úvodem několik slov o Českém národním korpusu (dále ČNK), z něhož materiálově vycházíme. Vzniká od roku 1992 a na jeho budování se podílí skupina odborníků ze zmíněných sedmi pracovišť na FF UK, MFF UK, FF MU, FI MU a ÚJČ ČAV. Od podzimu roku 1995 koordinuje práci na ČNK samostatný Ústav ČNK na FF UK v Praze. V současné době zahrnuje ČNK synchronní subkorpus čítající zhruba 70 milionů slovních tvarů, který má být v průběhu r.1998 rozšířen na 100 milionů tvarů. ČNK je dostupný na Internetu (http://ucnk.ff.cuni.cz/cnc). Dále vzniká diachronní subkorpus ve formě vzorků čítajících asi 950 tisíc slovních tvarů a mluvený synchronní subkorpus mluvčích z Prahy zahrnující zhruba 700 tisíc slovních tvarů. Komplementárně k němu se

na FF MU buduje vzorový subkorpus mluvené češtiny (asi 500 tisíc slovních tvarů) zachycující promluvy mluvčích narozených v městě Brně. Tento korpus je nyní dostupný ve formě přepisu pořízených nahrávek, v němž jsou jednotlivým slovním formám přiřazeny gramatické značky, má tedy vedle své základní podoby i formu částečně značkovanou. Jeho podstatná část tvořená kvalitně nahranými úseky byla nedávno převedena do digitalizované podoby na CD a bude takto posléze přístupná pro výzkumné účely. (Značkování přepsaného mluveného korpusu je popsáno v diplomové práci D. Hlaváčkové (Hlaváčková, 1998). Digitalizaci mluvených textů z uvedeného subkorpusu a jejich vypálení na CD provedli studenti FI MU Zrůstek a Vydržal.)

Jak jsme již řekli, spoluprací FF MU a FI MU vznikl (a dále se rozšiřuje) uvedený značkovaný synchronní subkorpus DESAM (rovněž přístupný na internetové adrese: http://www fi.muni.cz/~pary/korp, viz též Pala, Rychlý, Smrž, 1997) čítající něco přes milion gramaticky značkovaných slovních tvarů. Je sestaven z textů novinových (LN, MF DNES), populárněvědných (časopis Vesmír), ekonomicko-publicistických (Českomoravský profit) a také textů odborných (časopis Chip a uživatelský manuál k programu PowerPoint), které pocházejí z období 1992-96.

Při budování ČNK se zatím postupuje tak, že nejdříve se do korpusu začleňují nejsnáze dostupné soubory textů, což jsou z pochopitelných důvodů texty publicistické (jsou dnes k dispozici v podobě počítačových souborů pro sazbu, na CD-ROM nebo je lze stahovat z WWW-stránek) a texty odborné. Postupně přibývají i texty z dalších stylových oblastí tak, aby byla zajištěna co nejvyšší reprezentativnost ČNK.

2. Značkovaný korpus DESAM

Jestliže jazykový korpus chápeme jako vnitřně strukturovaný, jednotně zpracovaný a rozsáhlý soubor elektronicky uložených jazykových dat (textů) vytvořený obvykle pro určité cíle, pak ve značkovaném korpusu je navíc každému slovnímu tvaru přiřazena gramatická značka. Rozsah informace, kterou gramatická značka nese, může být v různých korpusech různý a je zjevně závislý na lingvistických teoriích, z nichž vycházejí ti, kdo značkování provádějí. Z dosavadních zkušeností korpusové lingvistiky ovšem vyplývá (Leech, 1993), že čím jsou lingvistické teorie, o něž se autoři značek opírají, transparentnější a neutrálnější ve vztahu k často soupeřícím lingvistickým školám, tím je korpus použitelnější pro co nejširší okruh zkoumání. Při budování značkovaného subkorpusu DESAM jsme se opírali o soubor značek, který je podle našeho názoru v dobré shodě se současným standardem gramatické teorie u nás (Havránek, Jedlička, 1981, Petr a kol., 1986).

2.1 Tvorba značkovaného korpusu

Nejprve připomeňme, že prvním plně značkovaným korpusem u nás byl korpus textů věcného stylu (dále VS) vytvořený pod vedením M.Těšitelové v oddělení matematické a kvantitativní lingvistiky v ÚJČ ČSAV na přelomu 70. a 80. let. Byl vytvořen manuálně, čítá 540 000 slovních tvarů a řadu výsledků

z jeho zpracování najde čtenář v práci M. Těšitelové (1985). V tomto textu údaje z VS v porovnávacích tabulkách neuvádíme — jednak jsme korpus VS neměli k dispozici a jednak by se tím náš text neúměrně rozrostl.

Podstatou gramatického značkování je vložení jisté interpretující informace do existujícího korpusu psaného nebo mluveného jazyka formou zvoleného symbolického zápisu (Leech, 1993). Rozlišujeme tedy korpusový text samotný a interpretaci k němu přidanou. Cílem gramatického značkování pak je opatřit každý slovní tvar v aktuálním korpusu značkou (tagem), která symbolicky reprezentuje gramatické významy nesené daným tvarem. V korpusu DESAM pracujeme se značkami, které mají následující strukturu: jsou definovány jako posloupnosti dvojic typu atribut:hodnota, kde atribut (značí se malým písmenem) reprezentuje některou z možných gramatických kategorií a symbol (velké písmeno nebo číslice) pro hodnotu vyjadřuje aktuální hodnotu, jíž daná kategorie u daného tvaru nabývá. Např. slovnímu tvaru *politik* přiřadíme značku **k1gMnSc1** a zachycujeme jí skutečnost, že tvar *politik* patří slovnědruhově k substantivům (k=1), nese kategorii rodu, a to mužského životného (g=M), nachází se v singuláru (n=S) a lze jej spojit s kategorií pádu (c=1), která zde nabývá hodnoty 1 (nominativ). Ke značce u substantiv (ale nejen u nich) ještě patří i údaj o vzoru, podle něhož se daný tvar ohýbá. Ten může u tvaru *politik* vypadat např. takto: *pán* Ea (o vzorech viz níže). Pro nedostatek místa zde nebudeme uvádět výčet užívaných značek, poznamenejme jen, že celkem je těmito značkami (viz též Hajič, Hladká, 1996,1997) pokryto obvyklých 10 slovních druhů a všech 14 gramatických kategorií, s nimiž se standardně setkáváme v českých gramatikách (Havránek, Jedlička, 1981, Petr a kol., 1986). Na rozdíl od současných gramatik vzniká navíc v korpusových textech potřeba značkovat systematicky např.číselné výrazy jako data, telefonní čísla, čísla výrobků a také speciální typy zkratek pro názvy firem či různých druhů a verzí výrobků (*Peugeot 406, Intel 486* apod.). Celkem v korpusu DESAM pracujeme s 1665 značkami. K tomuto poměrně vysokému číslu se dospívá možnými kombinacemi slovních druhů včetně subklasifikací (např. u zájmen jich je 8, u číslovek 4, u adverbií 6) s gramatickými kategoriemi, které se s jednotlivými slovními druhy standardně pojí. Porovnání našeho souboru značek např. s podobným soubory pro angličtinu, které čítají nejvýše kolem 200 značek, znovu potvrzuje vyšší morfologickou strukturovanost a bohatost češtiny jako silně flektivního jazyka.

Jestliže je naším cílem přiřadit značky tohoto typu každému slovnímu tvaru v korpusu čítajícím v našem případě něco přes milion slovních tvarů, je evidentní, že takovou práci nelze zvládnout manuálně (v zájmu korektnosti: dovedeme si představit, že by se o to někdo mohl pokoušet, ale pravděpodobnost takového konání je nepochybně dost nízká). Jediným rozumným a proveditelným řešením je použít počítačů. Pro značkování popsaného typu musíme pro češtinu nejprve použít **morfologického analyzátoru** (alternativně lze mluvit o lemmatizátoru, jestliže takový program přiřazuje slovním tvarům v textu vedle slovního druhu a příslušných gramatických kategorií i jejich tvary základní (lemmata). Je-li

takový program specializován primárně na značkování, což platí zejména v případě angličtiny, mluvíme pak o značkovacích programech (taggers). U češtiny výstup získaný z morfologického analyzátoru není ovšem jednoznačný a musí tedy projít další fázi zpracování, v níž se provádí **zjednoznačnění** čili desambiguace.

Základní značkování subkorpusu DESAM bylo provedeno morfologickým analyzátorem LEMMA (Ševeček 1995-96, Osolsobě, 1996). Tento analyzátor (lemmatizátor) pracuje na základě strojového slovníku čítajícího asi 164 000 českých kmenů a dovede každému rozpoznanému slovnímu tvaru ve volném textu přiřadit odpovídající základní tvar, tj. jeho **lemma**, a jak jsme už řekli, i gramatickou značku nesoucí údaje o příslušných gramatických kategoriích. Program LEMMA dovede také každému českému nominálnímu tvaru (ale i verbálnímu a dalším) přiřadit jeho **deklinační vzor** (obecně jakýkoli ohýbací), a to díky tomu, že je v něm zabudován algoritmický popis celé české flexe založený na detailní klasifikaci ohýbacích vzorů (Osolsobě, 1996). Jen u substantiv se v něm pracuje s 380 vzory (u adjektiv je to 66, u sloves 220), to ovšem z hlediska počítačového zpracování není mnoho a řeší se tím systematicky a elegantně potíže s tzv. výjimkami.

2.2 **Desambiguace**

Jak jsme již naznačili, při vytváření značkovaného korpusu se musíme vyrovnat s jednou podstatnou skutečností, která spočívá v tom, že lemmatizátor přiřazuje asi 70 procentům analyzovaných tvarů více než jednu značku. Naším cílem ovšem je, aby značkování bylo jednoznačné, musíme proto tvary s morfologickou homonymií (její míra zjištěná v korpusu DESAM činí 4,81 značky na tvar) desambiguovat. To lze udělat buď manuálně, nebo raději pokud možno automaticky s použitím vhodných programových nástrojů. Dnes se užívá nejčastěji desambiguačních programů založených na statistických a pravděpodobnostních přístupech (Hajič, Hladká,1996) nebo se pracuje s tzv. částečnými analyzátory (Abney, 1996), jejichž jádrem jsou formální (nekontextová) pravidla popisující strukturu příslušných větných složek.

Značkování korpusu DESAM probíhalo ve dvou fázích:

a) vybrané textové soubory (v rozsahu asi 250 tis.slovních tvarů) byly zpracovány programem LEMMA a pak desambiguovány manuálně pomocí speciálního prohlížecího programu DESAMB vytvořeného k tomuto účelu (Puža, 1997). Tím vznikla (v první polovině r.1997) první (trénovací) verze korpusu DESAM1, která byla po příslušných opravách použita k vyhledání a sestavení formálních (v zásadě nekontextových) pravidel popisujících strukturu českých jmenných skupin a gramatickou shodu uvnitř nich.
b) Na těchto pravidlech je postaven částečný syntaktický analyzátor DES implementovaný v Prologu (Puža, 1997), který byl použit k odstranění přibližně 40 procent nejednoznačných tvarů uvnitř jmenných skupin. Zbytek (asi 30 procent) desambiguovali manuálně pomocí již zmíněného interaktivního

prohlížeče DESAMB studenti (pomocné vědecké síly) FF MU a FI MU. Druhé kolo značkování proběhlo během druhé poloviny r.1997 podobným způsobem, ovšem podíl manuální desambiguace se již, jak jsme naznačili, podařilo výrazně redukovat.

Výsledkem je značkovaný korpus DESAM čítající 1 026 733 slovních tvarů, který v současnosti slouží mj. jako trénovací a testovací soubor dat pro vytvoření další verze desambiguátoru založeného na formálních (nekontextových, přesněji DC) pravidlech. U něho již počítáme s tím, že míra jeho úspěšnosti se bude blížit 90 procentům, takže potřeba manuální desambiguace se omezí, jak předpokládáme, z největší části na opravu chyb a řešení sporných případů.

3. Zpracování deklinačních vzorů

Naznačili jsme, že výchozí značkování korpusu DESAM bylo provedeno automaticky programem LEMMA, v němž je zabudován formální a velmi detailní popis české flexe, a že s jeho pomocí jsme mohli přiřadit odpovídající vzor každému nominálnímu tvaru v korpusu. Naše koncepce vzorů a konkrétně nominální deklinace se dosti liší od tradiční, a proto pokládáme za vhodné se o ní stručně zmínit.

Tradičně se v českých mluvnicích a v souladu s nimi i ve školských učebnicích uvádí 14 substantivních vzorů české substantivní deklinace. Automatická morfologická analýza si však z mnoha důvodů vyžádala mnohem podrobnější klasifikaci deklinačních paradigmat (srov. Osolsobě, 1996). Základní strategie, o niž se opírá systém vzorů, se kterými LEMMA pracuje, vychází z toho, že vedle skutečných výjimek, k nimž řadíme slova ojedinělá typu *člověk* atp., existují celé skupiny slov, které tvoří formálně dobře definovatelné „podvzory". Vzory jsou pak pro program LEMMA definovány jako možné kombinace kmenů s koncovkami. Koncovky jsou definovány samostatně jako skupiny (množiny) koncovek, kdy každá koncovka nese navíc gramatické významy. Koncovky jsou rozděleny do koncovkových množin na základě dvou kritérií:

a) koncovky, které potenciálně způsobují alternaci kmene
b) koncovky, které potenciálně alternují

Koncovky, jež nelze na základě uvedených kritérií vytřídit z morfologické charakteristiky (množiny koncovek tvořících paradigma), tvoří jádrové koncovkové množiny, koncovky, které lze na základě uvedených kritérií vydělit, se řadí do periferněji ších množin. Posledním kritériem pro vydělení podmnožin koncovek českých substantiv je potřeba mít samostatné koncovkové množiny pro definici flexe cizích slov a výjimek.

Každý z 380 substantivních vzorů, s nimiž program LEMMA pracuje, je jednak definován vzorovým slovem, které pomáhá autorovi (uživateli programu) v orientaci např. při přiřazování nových kmenů ke vzorům, jednak kódem, z nějž lze vyčíst, v jakém je nový vzor vztahu ke klasickému vzoru a na základě jakých kritérií je definován.

Vedle slovně uvedených klasických vzorů *pán, hrad,..., stavení* tedy pracujeme s kódy tvořenými velkými písmeny nebo kombinacemi velkých a malých písmen. Aby si čtenář mohl udělat základní představu o kritériích, která byla použita k získání podrobné subklasifikace vedoucí ke zmíněným 380 substantivním vzorům, uvádíme níže aspoň hlavní z nich (podrobněji o tom viz Osolsobě, 1996):

R — alternace prefixu (pRefix)
S — alternace ve kmeni (Stem)
F — alternace ve finále (Final)
U — alternace v (kmenotvorné) příponě (sUffix)
E — alternace v koncovce (Ending)
I — substantivum cizího původu (Import)
N — jiný tvar nominativu (Nominativ)
Y — kolísání tvrdý — měkký vzor
P — pomnožné substantivum
X — nepravidelné (eXtra)
a,b,c — další různé varianty

Dodejme ještě, že program LEMMA (rádi bychom také poděkovali dr. P. Ševečkovi, který pro potřeby tohoto výzkumu modifikoval morfologický analyzátor LEMMA tak, aby pro každý zpracovaný slovní tvar z textu dával automaticky i jeho ohýbací vzor) obsahuje vedle pravidel deklinace morfologická a částečně i slovotvorná pravidla definující kombinovatelnost kmenů uložených ve slovníku s českými koncovkami definovanými v rámci popisu českých koncovkových množin (Osolsobě 1996). Díky tomu lze automaticky rozpoznávat a odvozovat deverbativa, deadjektiva, deadverbia a posesivní adjektiva (od životných substantiv). Celkově lze říci, že program LEMMA pokrývá odhadem nejméně 300 000 českých lemmat a je-li použit jako generátor českých tvarů, může jich vytvořit minimálně kolem 6 000 000. (Po implementační stránce je LEMMA k dispozici pro všechny u nás používané platformy, tj. pro operační systémy DOS, UNIX (LINUX) a MACINTOSH. V interaktivním režimu umožňuje uživateli klást dotazy na jednotlivé slovní tvary, jež program LEMMA analyzuje, a přiřadí jim základní tvar, slovní druh a gramatické významy podle příslušného slovního druhu.) Při vytváření korpusu DESAM bylo použito dávkového zpracování. Nerozpoznaná slova (řetězy znaků) — nejčastěji zkratky, číselné výrazy, slova cizího původu ponechává LEMMA neoznačena — je jich počet se pohybuje kolem 1,4%. Taková slova se pak dodatečně zpracovávají ručně.)

4. Frekvenční analýza českých substantivních vzorů

Není asi třeba dvakrát zdůrazňovat, že bez značkovaného korpusu by se frekvence českých substantivních vzorů daly získat jen velmi těžko — díky němu lze příslušné substantivní tvary vyhledávat v korpusu a získávat číselné údaje o vzorech automaticky. Aby si čtenář mohl učinit představu, uveďme, že vytřídění

potřebných údajů (viz níže) o frekvencích deklinačních vzorů proběhlo v časovém horizontu hodin.

V této souvislosti pokládáme za vhodné dotknout se některých obecných metodologických aspektů v lingvistice a porovnat omezenost klasických postupů založených na technice excerpce s technikami korpusové a počítačové lingvistiky. Právě zkoumání frekvence vzorů v textu je příkladem toho, že korpusová a spolu s ní i lingvistika počítačová poskytují metodologické nástroje, bez nichž by se např. právě frekvence vzorů prakticky nedaly sledovat v rozumném časovém horizontu a také — v neposlední řadě — s přijatelnými finančními náklady. Manuální zpracování textů v rozsahu jednoho miliónu slovních tvarů a zjišťování četností vzorů technikou excerpce na kartičkách, tj. klasickým způsobem, by nepochybně zabralo několik člověkoroků (odhadem nejméně 2-3, pokud by se dnes vůbec našel člověk, jakým byla např. dr.Marie Těšitelová, který by ještě byl ochoten se do takové práce pustit). Přitom nejde o výzkum, který by byl teoreticky zvlášť komplikovaný, jen je především nesmírně pracný. Máme-li k dispozici dostatečně velký gramaticky označkovaný korpus vytvořený technikami korpusové lingvistiky a lemmatizační a desambiguační programy, které jsou výsledkem postupů vypracovaných počítačovou lingvistikou, můžeme výchozí statistiku frekvencí vzorů v textu o rozsahu asi jednoho milionu tvarů získat během dvou-tří hodin a celou práci na podkladové statistice vzorů tak provést řádově v časovém horizontu dnů. Statistiku slovních druhů, kterou uvádíme níže, lze pak získat doslova během několika minut.

Výsledky získané korpusovými postupy v dohledné době ovlivní ovšem nejenom samu lingvistiku, ale i současné počítačové zpracování přirozeného jazyka — již teď na základě korpusových dat vznikají nové a přesnější elektronické slovníky a robustní počítačové gramatiky. Korpusy jsou dnes v jazykovědě východiskem pro realistický **základní výzkum** ve formě relativně blízké přírodním vědám. Tento výzkum nepochybně povede k postupným úpravám a doplňování existujících gramatik a jazykových příruček a v blízké budoucnosti též k novému velkému (akademickému) výkladovému slovníku současné češtiny.

5. Výchozí údaje pro FSČ a DESAM

Na tomto místě nejprve připomeňme, že FSČ vznikal, jak známo, od 40.let, pracně a zdlouhavě pomocí excerpce a rozsáhlých kartoték. Dokonce ani po svém dokončení v r.1953 se díky nepříznivému politickému klimatu nedočkal okamžitého vydání a byl vytištěn až v roce 1961. Materiál FS obsahuje celkem 1 623 527 dokladů (slovních tvarů) pocházejících z různých stylových okruhů, tj.z beletrie, poezie, dramatu, mluvených projevů, literatury pro mládež, publicistiky, vědecké a odborné literatury. Z tohoto hlediska je materiál FSČ reprezentativnější než DESAM, který obsahuje jen publicistické a odborné texty, ovšem z hlediska sourodosti pokládáme DESAM za přirozenější a spolehlivější, protože jde o textový korpus, tedy o text v jeho přirozené podobě, a nikoli o excerpta.

Tab.1 Výchozí údaje

	FSČ	DESAM
počet dokumentů	75	3056
všechny slovní tvary (tokens)	1 623 526	1 026 733
různé slovní tvary (types)	—	132 447
lemmata	54 486	34 606
type/token ratio	—	7,75
hapax legomena (lemmata)	20 467	11 759
hapax legomena (tvary)	—	67 059

Komentář k tab.1

Základní porovnání v tab.1 ukazuje, že některé údaje pro FSČ chybí, např. poměr type/token, jehož hodnota získaná z DESAM dobře odráží vysoce flektivní povahu češtiny. Abychom je získali, museli bychom mít k dispozici originální kartotéku FSČ, — takový pokus by ovšem vedl ke klasickým potížím s bariérou manuálního zpracování. Jistý rozdíl mezi FSČ a DESAM je též v chápání dokumentů, ve FSČ se jimi obvykle rozumí knihy, části knih nebo čísel novin, v DESAM se za dokumenty pokládají jednotlivé novinové nebo časopisecké články, případně kapitoly (u počítačových manuálů). Je třeba konstatovat, že DESAM je sice rozsahem menší a v tomto ohledu i do jisté míry předběžný, ovšem za jeho přednost je třeba pokládat to, že je značkován: díky tomu a počítačovému uložení je opakovaně přístupný a použitelný i pro řadu dalších sond.

5.1 **Porovnání frekvence substantivních vzorů ve FS a DESAM**

Následující tabulka 2 obsahuje porovnání frekvence českých substantivních vzorů ve FSČ a DESAM (v procentech) a představuje tak jeden z hlavních výsledků, k nimž jsme na základě údajů z FSČ a korpusu DESAM dospěli.

Tab.2

vzory	FS %	DESAM %
pán:	8,9	7,3
hrad:	25,13	27,7
muž:	4,0	3,5
soudce:	0,4	0,5
stroj:	2,7	2,6
předseda:	1,10	0,7
celkem	**42,24**	**42,3**
feminina		
žena	22,0	22,4
nůše/růže:	6,21	8,2
píseň:	2,24	1,7
kost:	7,02	5,0
paní:	—	0,03
celkem	**37,47**	**37,33**

vzory	FS	%	DESAM	%
neutra				
město:		5,61		6,2
moře:		0,73		0,4
kuře:		0,36		0,3
stavení:		11,82		9,0
příjmení:		—		1,7
celkem		**18,52**		**17,6**
nesklonná		0,44		0,3
adjektivní		1,28		1,1
nezařazeno				1,4
celkem		**100**		**100**

Celkem překvapivě se ukazuje, že přes poměrně značný časový odstup mezi dobou vzniku FSČ a DESAM jsou rozdíly ve frekvencích deklinačních vzorů velmi malé. Neprojevuje se tu nijak výrazně ani větší rozsah FSČ, ani jeho větší stylová rozrůzněnost. Podle našeho názoru uvedené hodnoty celkem jednoznačně vypovídají o tom, že systém české substantivní deklinace ve spisovném jazyce je dlouhodobě stabilní a že v průběhu časového úseku asi 37 let nevykazuje výraznější změny. V získané statistice lze upozornit na vyšší četnost vzoru *pán* ve FSČ v porovnání s DESAM a obrácený poměr četností u vzoru *hrad*. Tuto skutečnost si vysvětlujeme vyšší stylovou rozrůzněností zdrojů FSČ, konkrétně přítomností uměleckých textů, v nichž lze očekávat vyšší výskyt životných substantiv. Podobně u feminin nenajdeme mezi frekvencemi vzorů z FSČ a DESAM větší rozdíly, snad jen u vzoru *kost* pozorujeme vyšší četnost ve FSP, kterou lze též připsat na vrub vyšší stylové různorodosti FSČ; může být způsobena vyšším výskytem abstrakt zakončených na *-ost*. U neuter lze v DESAM pozorovat poněkud vyšší četnost vzoru *město,* ale opravdu zajímavý by mohl být rozdíl mezi oběma zdroji u vzoru *stavení*, kde se rozdíl blíží třem procentům. I zde by se dalo uvažovat o stylových důvodech, FSČ v každém případě obsahuje jak umělecké, tak i odborné texty, což by mohlo vysvětlovat pozorovaný rozdíl. Proti tomu ovšem stojí fakt, že v DESAM se pracuje se samostatným vzorem *příjmení*, který má četnost 1,7%, ale může být (a ve FSČ jistě je) zahrnut pod vzor *stavení*. Pak i u neuter se potvrzuje vzácná vyrovnanost mezi vzory ve FSČ a DESAM.

Údaje v tab. 2 poskytují jasný obraz též o produktivitě jednotlivých vzorů. Tento obraz je ve shodě s dřívějšími pozorováními: u maskulin jsou nejproduktivnější *pán, hrad a muž*, u feminin jsou to *žena, růže, kost* a u neuter sem patří *město* a *stavení.*

Rádi bychom však upozornili, že DESAM nabízí mnohem detailnější pohledy na vzory a jejich subklasifikaci, což můžeme demonstrovat např. na vzoru *stavení*: díky výše zmíněnému podrobnému popisu vzorů můžeme tu rozlišit, kolik tento vzor zahrnuje deverbativ typu *braní, bytí, chování, chápání, hledání, pro-*

šení, nesení aj. a kolik substantiv typu *bezpečí, doživotí, dříví, zrní, hornictví, manželství.* Prvních je v DESAM 18280, tj.5,8% z celkových 10,6%, druhých potom 14000, tj. 4,8% z celkového počtu. Dodejme ještě, že jako nejproduktivnější zdroj deverbativ se u slovesných vzorů jeví vzor *prosí* (8767 případů) a podle očekávání následují vzory *dělá* (3231), *kupuje* (2643), *tiskne* (659) a *nese* (589). Pochopitelně, lze očekávat námitku, že není vždy možné jednoznačně rozhodnout, kdy dané substantivum je nepochybným deverbativem a kdy se má již řadit do druhé skupiny: ovšem celkem bezpečným vodítkem tu jsou některé sufixy, např. *-ctví, -ství.* Jde spíše o slovotvorný problém, který se tu nepokoušíme řešit, poznamenáváme však, že DESAM může v případě potřeby poskytnout rozsáhlé materiálové východisko pro jeho úplnější řešení. Podobně bychom se mohli detailně věnovat podvzorům v souvislosti s jednotlivými typy alternací, ale to jsou náměty pro samostatné dílčí studie.

Porovnání prvních 100 nejčetnějších substantiv ve FS a DESAM
Tab. 3

FS	abs.č.	DESAM	abs.č.
pán/pan	3447	rok	2654
život	2736	Praha	1958
člověk	2705	člověk	1646
práce	2486	auto	1361
ruka	2440	doba	1341
den	2287	cena	1165
zem/ě	2078	strana	1143
hlava	1802	koruna	1063
oko	1766	země	1061
10cesta	1665	společnost	1008
lid	1568	případ	982
léto	1548	práce	971
strana	1483	zákon	952
žena	1433	den	943
slovo	1416	místo	936
dítě	1405	vláda	891
síla	1316	systém	887
místo	1217	svět	827
rok	1207	možnost	799
20město	1156	většina	770
škola	1129	trh	763
národ	1110	podnik	751
muž	1098	problém	737
voda	1068	část	730
tvář	1000	program	719
hlas	967	skupina	715
bůh	945	republika	679

FS	abs.č.	DESAM	abs.č.
otázka	930	procento	672
noc	925	město	652
30Praha	898	ministr	622
srdce	889	Evropa	615
hodnota	866	čas	605
otec	849	otázka	603
pravda	849	výsledek	603
les	798	změna	599
matka	786	řada	584
případ	783	situace	577
hodina	764	oblast	577
tělo	754	služba	560
40láska	750	smlouva	556
smrt	731	prezident	554
sen	731	způsob	544
plán	727	informace	536
boj	704	počet	533
motor	701	kontakt	531
paní	689	snímek	527
dveře	688	člen	522
vývoj	686	hodina	518
republika	678	ministerstvo	514
50král	668	dítě	510
Právo	666	věc	510
Václav	660	voda	505
vláda	647	činnost	504
dělník	643	zájem	501
noha	638	objekt	500
část	631	muž	499
úkol	619	volba	488
dům	596	cesta	487
zákon	584	jednání	484
60válka	582	měsíc	482
kůň	581	fax	482
vzduch	575	úřad	479
jméno	575	návrh	477
smysl	574	funkce	477
příklad	569	podmínka	476
krok	555	slovo	473
třída	555	výroba	472
pole	547	dům	468
zboží	545	peníze	467
70forma	545	žena	444
řeč	538	hráč	444

FS	abs.č.	DESAM	abs.č.
poměr	534	ředitel	439
řada	531	předseda	436
světlo	528	týden	435
vůz	526	zařízení	435
chlapec	523	prostředí	431
duše	521	stav	419
stát	517	poslanec	417
dílo	516	hodnota	412
80stůl	510	osoba	410
stav	510	pracovník	408
stupeň	504	Jan	407
skutečnost	504	nabídka	405
Němec	503	počítač	403
ulice	503	typ	402
pohled	501	škola	399
píseň	501	milio/ón	396
hoch	501	náklad	396
tatínek	501	lékař	395
voják	500	názor	394
90způsob	500	podnikatel	393
krev	497	důvod	382
počet	493	zápas	380
Blažena	492	úroveň	378
myšlenka	487	daň	376
základ	486	tým	374
maminka	485	banka	374
duch	477	základ	369
bratr	473	bod	363
kníže	470	politika	362
100 syn	465	text	357

Časový odstup je tedy zhruba 37 let, proto lze očekávat, že údaje o současné češtině se budou od údajů FSČ zřetelně lišit. Je potřeba vzít v úvahu rozdílnou situaci politickou a ekonomickou a také výrazné změny v oblasti vědy a techniky. Rozdíly jsou dobře patrné v tab.3, která, jak lze vidět, obsahuje prvních 100 nejčetnějších substantiv ve FSČ a korpusu DESAM.

Nejprve se podívejme na substantiva, která se objevují shodně jak ve FSČ, tak i v DESAM: *člověk* (3,3), *práce* (4,12), *země* (7,9), *den* (6,14), *strana* (13,7), *rok* (19,1), *Praha* (30,2), *republika* (49,27) (čísla v závorkách uvádějí pořadí neboli rank v seznamu nejčetnějších substantiv: je-li u substantiva dvojice čísel, první platí pro FSPČ a druhé pro DESAM). Vedle ideologicky neutrálních substantiv jako *člověk, den, země, rok* stojí za povšimnutí *strana* — jeho vysoká frekvence ve FSČ je nepochybně způsobena tím, že odkazuje k tehdejší komunistické straně, která byla pilířem totalitního režimu. V DESAM je vysoká frek-

vence tohoto substantiva jednak odrazem toho, že současný politický systém je založen na pluralitním systému politických stran a jednak i projevem prozaické skutečnosti, že v novinových textech se při označování stránek užívá pravidelně právě tohoto substantiva (*strana 1, 2,...*). Poznamenejme, že zajímavý rozdíl lze pozorovat u substantiva *Praha,* které se v DESAM vyskytuje na druhém místě s četností 1958, zatímco ve FSČ na místě třicátém s četností 898. Mohlo by to vést k závěru, že současný pragocentrismus je silnější než ten dřívější, ale v zájmu korektnosti je nutno vzít v úvahu i to, že DESAM je tvořen z největší části novinovými texty, které opravdu často představují Prahu jako pupek světa.

Zajímavé jsou skupiny substantiv, v nichž se oba seznamy liší: nejprve uveďme skupinu substantiv, která najdeme v první stovce ve FSČ, ale nikoli v DESAM. Patří sem např. substantiva *lid* (11), *národ* (22), *Právo* (51), *dělník* (54), *válka* (60), *boj* (60), *třída* (67), *zboží* (69), *Němec* (84), která podle našeho názoru jednoznačně odrážejí převažující ideologickou orientaci v období lidově demokratického zřízení, kdy FSČ vznikal. Poněkud stranou je substantivum *zboží,* které jediné z uvedených patří do oblasti ekonomické: zdá se, že přes svůj deklarovaný význam nebyla ekonomika v předchozím režimu tak docela v popředí zájmu.

V DESAM je protějškem následující skupina substantiv: *zákon* (13), *vláda* (16), *svět* (18), *ministr* (30), *Evropa* (31), *prezident* (41), *úřad* (62); již tato malá skupina svědčí jasně o tom, k jak velkým politickým změnám u nás došlo a stále dochází. Do očí bije fakt, že zatímco substantivum *zákon* má v DESAM vysoké pořadí — 13, ve FSČ se v první stovce vůbec nevyskytuje.

Dále v DESAM najdeme mezi prvním stem nejčetnějších substantiv dvě zajímavé skupiny, první z nich bychom mohli charakterizovat přívlastkem „technická": *auto* (4), *systém* (17), *program* (25), *informace* (43), *fax* (61), *počítač* (83). Z nich třeba *fax* a *počítač* nenajdeme ve FSČ vůbec, což je pochopitelně způsobeno tím, že před 37 lety se nám o současném bouřlivém rozvoji informačních technologií mohlo jen zdát. K tomu dodejme, že v DESAM má poměrně vysokou frekvenci (absolutní — 111) též substantivum *myš,* což je docela pěkný příklad metaforického použití substantiva označujícího normálně zvíře (v odborných i publicistických textech je to synonymum pro tzv. polohovací zařízení, které je standardní součástí počítačového vybavení). Příznačná je i vysoká četnost substantiva *auto* (4) v korpusu DESAM: prozrazuje současnou posedlost tímto často smrtonosným symbolem prosperity a úspěšnosti. Ve FSČ toto substantivum v první stovce nejčetnějších podstatných jmen nenajdeme, vyskytuje se tam jen substantivum *vůz* s rankem 75.

Druhá skupina se dá označit atributem „ekonomická": *cena* (6), *koruna* (8), *trh* (21), *podnik* (22), *smlouva* (40), *nabídka* (83), *náklad* (88), *podnikatel* (90), — odrážejí změněnou ekonomickou orientaci typickou pro současnost. Ani tato skupina substantiv nemá přímý protějšek ve FSČ, ale i zde je přirozené vysvětlení nasnadě: v novinových textech tvořících DESAM můžeme oprávněně očekávat vyšší frekvenci těchto substantiv.

Konečně bychom rádi upozornili na jednu skupinu substantiv, která najdeme v první stovce ve FSČ, ale nikoli v DESAM: *hlava* (8), *oko* (9), *srdce* (31) (části

těla) a *bůh* (27) (psáno ovšem po ateisticku s malým b), *pravda* (34), *láska* (40), *smrt* (41), *sen* (42), *duše* (77), *duch* (97). Bylo by asi ukvapené vyvozovat z této skutečnosti jakékoli závěry o případných změnách v našem morálním klimatu v současném období, spíše je to, domníváme se, důsledek toho, že FSČ na rozdíl od DESAM obsahuje texty z krásné literatury, poezie a dramat. Ve shodě s M.Těšitelovou (Těšitelová, 1985) můžeme konstatovat, že vysokou frekvencí se i v současných publicistických textech vyznačují substantiva s významem časovým a prostorovým: v první dvacítce jich je 7, tj.35% — *rok, Praha, doba, země, den, místo, svět.* Zvlášť je třeba upozornit na substantivum *den* (a také *týden*), které si přes svou tvarovou nepravidelnost (pozůstatek kmenové deklinace) udržuje své postavení mezi ostatními vysoce frekventovanými substantivy.

Podíváme-li se na distribuci rodů u substantiv v první dvacítce DESAM, zjistíme jednoznačnou převahu feminin — je jich 11 (55%): *doba, cena,koruna, strana, Praha, země, společnost, práce, vláda, možnost*; za nimi následují maskulina v počtu 7 (35%): *rok, člověk, případ, zákon, den, systém, svět*, a na poslední místě jsou neutra: *auto, místo* (10 %). Ve FSČ je distribuce rodů v první dvacítce poněkud vyrovnanější: maskulina mají 40% a feminina i neutra mají po 30%. Pokud jde o životnost, je mezi prvními dvaceti substantivy v DESAM jen jedno životné substantivum *člověk*, pokud bychom k nim nepřidali ještě femininum *vláda*, které by se s jistou licencí dalo považovat za životné. Ve FSČ lze k životným podobně řadit 5 substantiv: *pán/pan, člověk, lid, žena, dítě*, což opět svědčí o vyrovnanějším zastoupení vyplývajícím nepochybně z většího rozsahu FSČ.

Jednou z výrazných tendencí je, že proti FSČ se v DESAM objevuje řada substantiv cizího původu s dosti vysokou frekvencí. Jen namátkou uveďme z první stovky: *kontakt, informace, fax, funkce, tým, text.* Další se najdou v druhé stovce, např. *fond* (294), *parlament* (289), *organizace* (286), *centrum* (257), *projekt* (239), *prezentace* (232).

Z korpusu DESAM se také získali informace o situaci v současných textech (1992-96) ve vztahu k novým pravidlům českého pravopisu (Hlavsa et al, 1993). Podíváme-li se na často probírané a tedy nejproblematičtější případy dvojího způsobu psaní slov přejatých, dostáváme následující obraz (číslo v závorce udává absolutní četnost jednotlivých dubletních variant v korpusu DESAM):

kurs (104) : *kurz* (119)
president (3) : *prezident* (580)
impuls (12) : *impulz* (1)
diskuse (104) : *diskuze* (3)
milion (615) : *milión* (265)
filosofie (9) : *filozofie* (88)
universita (3) : *univerzita* (142)

Po 5 letech existence „nových" pravidel českého pravopisu (Hlavsa et al, 1993) korpusové texty ukazují, že buď je poměr obou možných variant zcela vyrovnaný, jak je tomu u dvojice *kurs/z*, nebo naopak celkem jednoznačně vychýlený na jednu či druhou stranu (např. *prez/sident*). Nezdá se tedy, že by-

chom zatím mohli pozorovat souvislou a výraznější tendenci, která by naznačovala, kam by se snad mohl ubírat další vývoj. Současná situace spíše naznačuje, že v současných textech vzrostla míra „rozkolísanosti" nebo dokonce „zmatku" a „libovůle".

Korpusové texty také naznačují, že tzv."konservativní" psaní se celkem zřetelně preferuje a prosazuje v odborných textech (dokládají to texty z čas. Vesmír), dokonce se zdá, že se v nich nyní začíná uplatňovat i tam, kde tomu tak dříve nebylo (viz např. dvojice: *filósofie : filozofie,* ev. i další jako třeba *milion : milión*). Data z korpusových textů a zdravý rozum nás vedou k závěru, že současný stav rozkolísanosti není věci na prospěch a celkem zřetelně volá po nějakém racionálnějším řešení. Kloníme se k názoru, že rozumné a proveditelné řešení by se mělo osvobodit od současných inkonzistencí/inkonsistencí vyvolaných nedůslednými úvahami o potřebě další fonetizace i tak již dost fonetického českého pravopisu — viz např. dvojici *kurz/s* proti *dub,* a mělo by tedy vést ve směru spíše mírně konservativním.

6. Porovnání frekvencí slovních druhů ve FSČ a DESAM

Závěrem nabízíme tabulku 4, která obsahuje údaje o četnostech slovních druhů ve FSČ a DESAM (v procentech).

Tab. 4

slovní druh	**FSČ%**	**DESAM%**
substantiva,k1	27,77	33,75
adjektiva,k2	11,16	13,25
zájmena,k3	10,91	8,34
číslovky,k4	1,61	1,75
slovesa,k5	18,15	15,22
adverbia,k6	10,29	5,7
předložky,k7	10,12	12,0
spojky,k8	9,78	5,93
partikule,k9	—	2,62
citoslovce,k0	0,21	0,0008
zkratky,kX	—	1,44
celkem	**100**	**100**

Porovnání hodnot četností jednotlivých slovních druhů v tab. 4 ukazuje, že u substantiv se zřetelně projevuje a potvrzuje očekávaný rozdíl vyplývající z širšího stylového záběru FSČ proti DESAM. Vyšší četnost substantiv v DESAM a podobně i v korpusu věcného stylu (dále VS, Těšitelová, 1985) jde jistě na vrub publicistických a odborných textů tvořících DESAM a VS. Rozdíl je patrný též u sloves, domníváme se, že jejich naopak vyšší četnost ve FSČ je důsledkem skutečnosti, že FSČ zahrnuje umělecké texty, vyznačující se vyšší dynamičností, jež má svůj formální odraz ve vyšší četnosti slovesných lemmat.

Celkově dobrou shodu mezi FSČ a DESAM lze pozorovat u adjektiv (11,16–13,25) a číslovek (1,61–1,75). Jistý rozdíl u předložek (10,12 - 12,0) si vysvět-

lujeme tím, že DESAM obsahuje stylově specializovanější texty než FSČ, proto nepřekvapuje, že v DESAM je frekvence předložek vyšší.

Nejpřekvapivější rozdíly mezi FSČ a DESAM nacházíme u částic a zkratek. FSČ neuvádí pro částice žádnou hodnotu, v korpusu VS pak najdeme nízkou hodnotu 0,37. Jistě nejde o náhodu ani o opomenutí a stejně tak lze sotva akceptovat případné vysvětlení, že by se v excerptech FSČ částice vůbec nevyskytovaly. Spíše tu jde, jak se domníváme, o důsledek jistého teoretického postoje, díky němuž byly částice ve FSČ v zásadě ponechány stranou. Rovněž se ve FSČ a VS nepočítá se zkratkami — což plyne z faktu, že ve standardních gramatikách se buď o zkratkách nemluví vůbec, nebo jen zcela okrajově. Přitom je zřejmé, že představují stejně plnoprávnou skupinu jazykových výrazů, jako jsou třeba substantiva, přesněji řečeno, většina zkratek má jednoznačně substantivní povahu a zkratky syntakticky představují i docela složité nominální skupiny (a významově pak komplexní pojmenování), jejichž ignorování způsobuje, že realistická analýza textu potom nemůže být úplná. I zde korpusové přístupy naznačují, kterým jevům je v blízké budoucnosti potřeba věnovat soustavnější pozornost. Výše jsme např. uvedli pořadí substantiva *koruna* a *republika*. Při úplnější analýze bychom v této souvislosti museli vzít v úvahu i zkratky ČR a *Kč*, které se v DESAM vyskytují s absolutní četností 454 a 350. Podobně to platí o řadě dalších zkratek jako např. USA (454), ODS (172), SR (109), OSN (109), ČSSD (88), ODA (80) a dalších. Jejich samostatná frekvenční analýza opřená o korpusová data bude velmi potřebná i pro předpokládané standardní automatické zpracování volného textu.

7. Závěr

Na základě frekvenční analýzy substantivních vzorů v češtině jsme konstatovali celkovou stabilitu systému české deklinace. Její výsledky ve vztahu k frekvenci jednotlivých podvzorů užitých automatickým morfologickým analyzátorem LEMMA budou sloužit jako východisko pro návrh algoritmu pro poloautomatizované zařazování slov, která analyzátor LEMMA neidentifikuje, do slovníku kmenů (srov. Osolsobě, 1996). Navíc nám materiál získaný analýzou části ČNK otevírá cestu k celé řadě dalších úvah jak v oblasti slovotvorby, tak v oblasti formální morfologie a samozřejmě i významosloví. Kromě analýzy vzorů jsme získali konfrontaci první stovky substantiv nejčastěji zastoupených ve FSČ a DESAM i pohled na hlavní tendence ve vývoji slovní zásoby, jak se projevují v současných textech publicistického a odborného stylu. Konečně jsme mohli nabídnout i základní porovnání frekvencí slovních druhů. Úplně na závěr bychom rádi konstatovali, že s růstem velikosti ČNK a jeho značkovaných subkorpusů bude možno v blízké budoucnosti dospět k řadě nových a zajímavých výsledků — obrazně řečeno, korpusová data z ČNK umožní otevřít doslova „továrnu" pro lingvisty.

FREQUENCIES OF INFLECTIONAL PATTERNS OF CZECH NOUNS

The presented article offers the comparison of the quantitative characteristics of Czech nouns, particularly, the frequencies of Czech inflectional patterns, which form the skeleton of the Czech declension system in contemporary Czech, with the older findings as presented in the Frequency Dictionary of Czech (1961, 1 623 536 word forms). These results have been obtained from the new electronic source: **Czech National Corpus**, particularly from its grammatically annotated subcorpus DESAM containing 1 026 733 word forms. All the presented results have been obtained by the techniques developed in the framework of corpus and computational linguistics (automatic lemmatization, partial automatic desambiguation, etc.).

The comparison shows that Czech declension system is very stable and that the span of 37 years has influenced the frequencies of Czech inflectional patterns only in small and non-distinctive details. The second important result offers comparative data from FSČ and DESAM for first 100 most frequent Czech nouns — here one can observe the basic tendencies and changes in Czech lexicon in the course of last 37 years. The third group of results follows from table 4 which compares the relative frequencies of the parts of speech in FSČ and DESAM.

The presented results obtained from ČNK and its subcorpora are just the beginning: the corpus data and techniques are opening the door for a new research and new and more precise findings about Czech language.

BIBLIOGRAFIE

1. HAVRÁNEK, B., JEDLIČKA, A.: Česká mluvnice, SPN, Praha 1981.
2. HAJIČ, J., HLADKÁ B., Probabilistic and rule based tagging of an inflective language — a comparison, Technical Report No.1, ÚFAL MFF UK, November 1996.
3. HAJIČ, J., HLADKÁ, B.: Tagging Inflective Languages: Prediction of morphological categories for a rich, structural tagset, *Technical Report TR-1997-04*, ÚFAL MFF UK, Praha
4. HLAVÁČKOVÁ, D.: Korpus mluvené češtiny, diplomová práce, Brno 1998.
5. HLAVSA, Z.a kol.: Pravidla českého pravopisu, Praha 1993.
6. JELÍNEK, J., BEČKA, J. V., TĚŠITELOVÁ, M.: Frekvence slov, slovních druhů a tvarů v českém jazyce, SPN, Praha 1961.
7. LAMPRECHT A., ŠLOSAR D., BAUER J., Historická mluvnice, češtiny, SPN, Praha 1986.
8. LAMPRECHT A., ŠLOSAR D., BAUER J.: Vývoj mluvnického systému českého jazyka, SPN, Praha 1970.
9. LEECH, G.: Corpus Annotation Schemes, in *Literary and Linguistic Computing*, Vol.8, No.4, 1993, 275–281.
10. TĚŠITELOVÁ, M., a kol.: Kvantitativní charakteristiky současné češtiny, řada *Studie a práce lingvistické*, Academia, Praha 1985.
11. TĚŠITELOVÁ , M. a kol.: O češtině v číslech, *Malá jazyková knižnice*, Academia, Praha, 1987.
12. OSOLSOBĚ, K.: Algoritmický popis české formální morfologie a strojový slovník češtiny, disertační práce, FF MU Brno 1996.
13. PALA K., Osolsobě K.: Základy počítačové lingvistiky, FF MU, Brno 1996.
14. PALA, K.: Korpusová lingvistika - informační technologie v lingvistice, Zpravodaj ÚVT MU, Brno 1996.
15. PALA, K., RYCHLÝ, P., SMRŽ, P.: DESAM — Annotated Corpus for Czech, *Proceedings of SOFSEM'97*, Springer Verlag, New York, Hamburg 1997.
16. PETR, J. a kol.: Mluvnice češtiny II., Academia, Praha 1986.
16. ŠEVEČEK, P.: Morfologický analyzátor (lemmatizátor) LEMMA, program v jazyce C, Brno 1995–96.

Klára Osolsobě
Ústav českého jazyka
Filosofická fakulta Masarykovy university
Arna Nováka 1
660 88 Brno
klara@ernest.phil.muni.cz

Karel Pala
Katedra informačních technologií
Fakulta informatiky Masarykovy university
Botanická 68a
602 00 Brno
pala@fi.muni.cz

Pavel Rychlý
Katedra informačních technologií
Fakulta informatiky Masarykovy university
Botanická 68a
602 00 Brno
pary@fi.muni.cz