K úkolům výzkumného záměru *Vytvoření databáze lexikální zásoby českého jazyka počátku 21. století*[1]

Albena Rangelova
Ústav pro jazyk český AV ČR, v. v. i.

On the Objectives of the Institutional Research Plan *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century*
This paper informs on the wider framework of the research plan *Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century*, where the central position is taken by LEXIKON 21 and PRALED. Further goals of the Department of Lexicography and Terminology of the Institute of the Czech Language ASCR, v. v. i. include consolidation of the existing material collections and descriptive databases in cooperation with the Department of the Data Electronisation: digitisation of the collections of excerption slips, creation of new excerption databases and digitised versions of dictionaries. The final objective is to present our results to both professional and general public.

Strategickým cílem výzkumného záměru *Vytvoření databáze lexikální zásoby českého jazyka počátku 21. století* (2005–2010), realizovaného v Ústavu pro jazyk český AV ČR, v. v. i. (dále jen ÚJČ), je **komplexní příprava na tvorbu moderního výkladového slovníku**. Na realizaci tohoto jedinečného projektu se podílí zejména lexikograficko-terminologické oddělení (LTO) ve spolupráci s úsekem pro elektronizaci dat. O jednotlivých aspektech centrálního proudu prací – o navržení softwarových nástrojů, o řadě koncepčních i realizačních otázek – pojednávají další příspěvky členů lexikograficko-terminologického oddělení v tomto sborníku. Zde chceme stručně informovat o **širším rámci** tohoto výzkumného záměru, který zahrnuje různorodé činnosti zaměřené zejména na vytváření materiálových, technických a personálních předpokladů pro lexikografickou práci, včetně prezentace lexikologického a lexikografického výzkumu pro širší veřejnost s využitím moderních informačních technologií.

Tento širší rámec směřuje k vytvoření **integrovaného databázového systému** obsahujícího řadu dílčích složek, které zatím existují buď izolovaně, anebo jen v tištěné podobě. V dlouhodobém výhledu budou dílčí celky postupně propojovány do jednotného vyhledávacího prostředí, které umožní efektivní vyhledávání jak v databázi jako celku, tak i v jednotlivých dílčích souborech. Zpřístupněná data

[1] Příspěvek vznikl v rámci výzkumného záměru ÚJČ AV ČR, v. v. i. *Vytvoření databáze lexikální zásoby českého jazyka počátku 21. století* (AV0Z90610521).

bude možné dále využívat pro různé lingvistické účely, např. pro další lexikografické zpracování i pro jakékoli obecně lingvistické studie v oblasti lexikonu.

Při plnění úkolů výzkumného záměru počítáme s využitím vědeckých výsledků řady **kooperujících pracovišť**. Rozsáhlá spolupráce byla navázána především s Centrem zpracování přirozeného jazyka FI MU v Brně. Dále spolupracujeme s Ústavem teoretické a komputační lingvistiky FF UK a s Ústavem formální a aplikované lingvistiky Matematicko-fyzikální fakulty UK. Zvláště významným partnerem je též Ústav Českého národního korpusu při FF UK, spravující rozsáhlé textové korpusy *SYN2000, SYN2005, SYN2006PUB* ad.

Stávající výzkumný záměr je v souladu s celkovou strategií ÚJČ budovat a postupně zpřístupňovat datovou základnu slovního bohatství českého jazyka tak, aby bylo možné jak její další rozšiřování, tak i optimální využití. Z pohledu strategických cílů pracoviště jde také o přípravu existujících primárních a sekundárních zdrojů lexikálního materiálu pro jejich další využití na nové technologické úrovni – konkrétně o skenování a popis lexikálních sbírek, digitalizaci slovníků, které na našem pracovišti vznikaly, o převedení našich elektronických sbírek na vyšší technologickou platformu ap. Cílem je zmíněné jednotné uživatelské prostředí spojující celou řadu dílčích databází, popisných i materiálových.

Ústřední postavení mezi popisnými databázemi bude mít nový, moderně pojatý lexikografický popis českého lexika, který vyžaduje rozsáhlou technickou a materiálovou přípravu. Od zahájení prací na výzkumném záměru se proto naším cílem stalo vytvoření **lexikografické pracovní stanice** specializované pro naše potřeby: jednak navržení programového vybavení pro popis lexikálních jednotek – program *Praled*[2] (srov. příspěvek J. Světlé v tomto sborníku) a pro práci s materiálem (*Pramat*, viz zde příspěvek Z. Opavské a B. Štěpánkové), budování a naplňování databáze s názvem Pralex, jednak propracování koncepčních otázek budoucího mnohoaspektového popisu slovní zásoby (slovníkové databáze s názvem *LEXIKON 21*, viz také v tomto sborníku příspěvky J. Světlé, M. Voborské, E. Birkhahnové a V. Chudomelové).

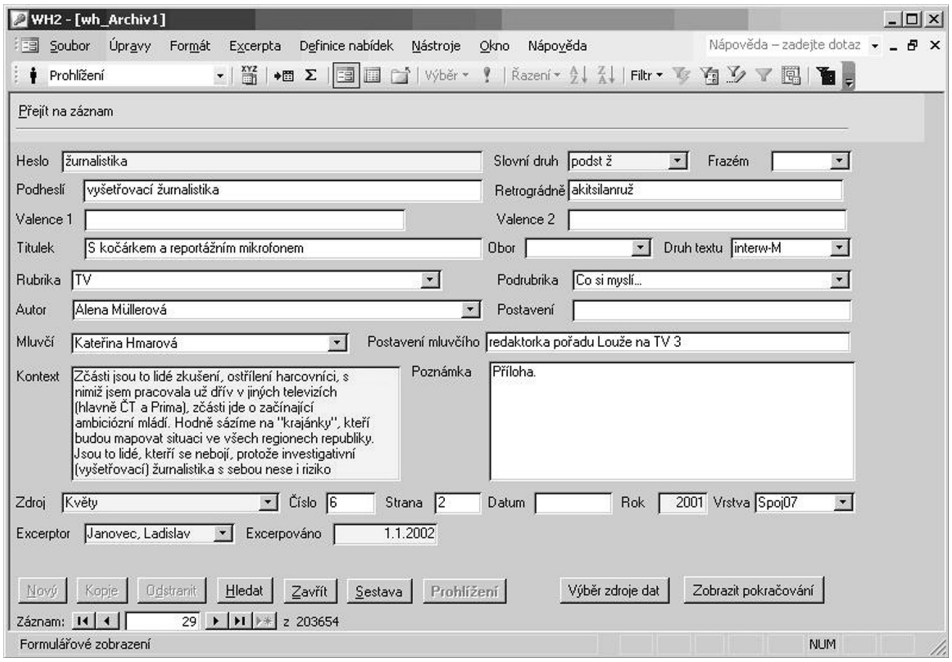
Dále usilujeme o vytvoření řady **slovníkových databází** představujících už publikované slovníky převedené do elektronické podoby. Již nyní jsou dostupné v různých aplikacích a v různém režimu následující díla: *Příruční slovník jazyka českého* (PSJČ), *Slovník spisovného jazyka českého* (SSJČ), *Slovník spisovné češtiny pro školu a veřejnost* (SSČ a SSČ-L), *Akademický slovník cizích slov* (ASCS, srov. též VSCS). Jako elektronicky uložené texty jsou k dispozici pro interní účely výzkumného týmu tyto publikace: *Retrográdní slovník současné češtiny* (na bázi Českého akademického korpusu), *Slovník slovesných, substantivních a adjektivních vazeb a spojení*, *Nová slova v češtině. Slovník neologizmů 1* (SN 1), Nová *slova v češtině. Slovník neologizmů 2* (SN 2), *Český jazykový atlas 1*, pracuje se na elek-

[2] Srov. též Pala – Horák – Rambousek – Rangelova 2007.

tronizaci díla *Slovesa pro praxi. Valenční slovník nejčastějších českých sloves*, výhledově se počítá s naskenováním publikace *Co v slovnících nenajdete. Novinky v současné slovní zásobě*. Při budování tohoto úseku databázového systému je třeba respektovat nejen autorská práva, ale též licenční práva a zájmy nakladatelů, proto zpřístupnění jednotlivých děl bude mít určitá omezení a diference. Bude sloužit zejména pro vyhledávání dílčích informací k vědeckým či jiným nekomerčním účelům.

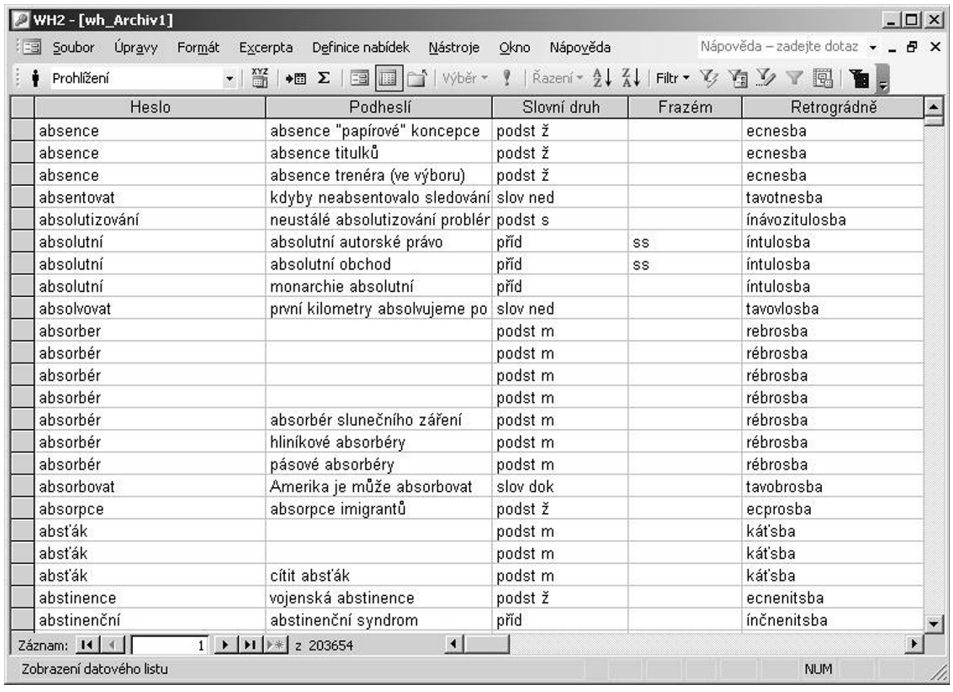
Materiálové databáze, kterými LTO ÚJČ disponuje, jsou dvojího druhu – část byla navržena a budována jako softwarové produkty s různým účelem, část vzniká postupnou elektronizací lexikálních sbírek pracoviště.

První kroky k elektronickému ukládání lexikálního materiálu byly v UJČ podniknuty již počátkem 90. let (první záznam je z r. 1991), kdy se excrepta ukládala do databázového programu Hesla v prostředí FOXBase, později FOXPro[3]. V rámci projektu *Popis nové slovní zásoby s využitím počítačové techniky* (1994–1996) byla založena neologická databáze LTO, která se postupně doplňovala novými daty v průběhu následných grantových projektů (*Systémotvorné procesy neologizmů v současné češtině*, 1998–2000, *Internacionalizmy v nové slovní zásobě češtiny*, 2001–2003). Její první část s pracovním názvem Archiv 1 obsahuje na 203 000 elektronicky uložených excerpt uložených v programu WinHesla2 (viz obr. 1, 2).



Obr. 1. Program WinHesla2 – excerpční karta

[3] Podrobněji viz Rangelova 1996.

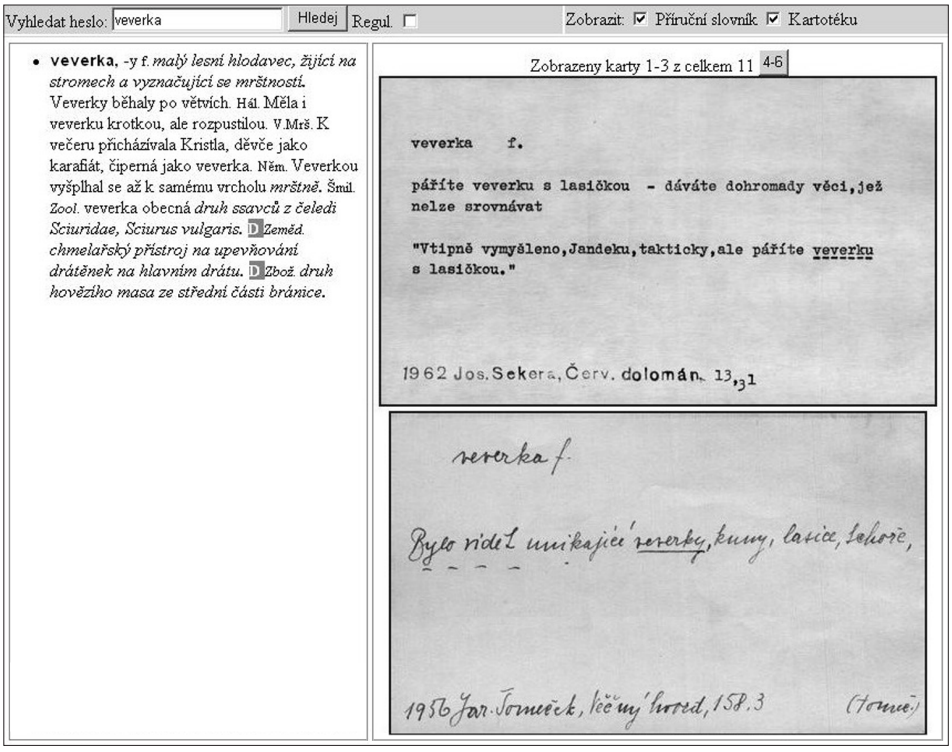


Heslo	Podheslí	Slovní druh	Frazém	Retrográdně
absence	absence "papírové" koncepce	podst ž		ecnesba
absence	absence titulků	podst ž		ecnesba
absence	absence trenéra (ve výboru)	podst ž		ecnesba
absentovat	kdyby neabsentovalo sledování	slov ned		tavotnesba
absolutizování	neustálé absolutizování problér	podst s		ínávozitulosba
absolutní	absolutní autorské právo	příd	ss	íntulosba
absolutní	absolutní obchod	příd	ss	íntulosba
absolutní	monarchie absolutní	příd		íntulosba
absolvovat	první kilometry absolvujeme po	slov ned		tavovlosba
absorber		podst m		rebrosba
absorbér		podst m		rébrosba
absorbér		podst m		rébrosba
absorbér		podst m		rébrosba
absorbér	absorbér slunečního záření	podst m		rébrosba
absorbér	hliníkové absorbéry	podst m		rébrosba
absorbér	pásové absorbéry	podst m		rébrosba
absorbovat	Amerika je může absorbovat	slov dok		tavobrosba
absorpce	absorpce imigrantů	podst ž		ecprosba
abstťák		podst m		káťsba
abstťák		podst m		káťsba
abstťák	cítit abstťák	podst m		káťsba
abstinence	vojenská abstinence	podst ž		ecnenitsba
abstinenční	abstinenční syndrom	příd		ínčnenitsba

Obr. 2. Program WinHesla2 – obrazovka hesláře

Tento jedinečný soubor neologického lexikálního materiálu dal vzniknout dvěma slovníkům neologizmů – SN 1 a SN 2 a sborníku statí *Neologizmy v dnešní češtině* (2005). V souladu s novými úkoly lexikograficko-terminologického oddělení byla od r. 2006 excerpce rozšířena na jevy synchronní dynamiky, nezachycené v dosavadních slovníkových dílech. Byla aktualizována též metodika práce – mnohem aktivněji se využívají elektronické textové archivy (Newton) a internetové zdroje. Specifickým úkolem je zabezpečení kompatibility databáze neologického materiálu s novým programovým vybavením pracoviště. Excerpční program WinHesla2 (programátor B. Lehečka, viz obr. 1 a 2) funguje v databázovém prostředí MS Access 2003, které má svá omezení, a proto se počítá s převedením dat a uživatelského rozhraní na flexibilnější platformu. Připravujeme též rozšíření excerpčních prací o odbornou slovní zásobu (výrazivo z různých oborů a oblastí lidské činnosti) v samostatné specializované databázi.

Obrovský význam má **databáze vytvořená na základě lexikálního archivu** ÚJČ, obsahující na 9,5 milionů lístkových excerpt všeobecné (novočeské) slovní zásoby. Jde o materiál neocenitelné hodnoty, díky němuž vznikly reprezentativní výkladové slovníky češtiny (PSJČ, SSJČ, SSČ). Plný soubor excerpt je již oskenován, vyhledávací programy jsou ve fázi testování a databáze (programátor M. Spousta) je postupně doplňována novými daty (snímky excerpčních lístků



Obr. 3. Slovo vyhledané současně v PSJČ [9] a v databázi lexikálního archivu

v elektronické podobě). Unikátní je propojení databáze lexikálního archivu s elektronickou podobou PSJČ (programátor P. Květoň), umožňující práci s oběma soubory zároveň (viz obr. 3).

Do elektronické podoby jsou převedeny rovněž menší, svébytné sbírky lingvistické a technické terminologie (na 300 000 elektronicky dokumentovaných a komentovaných excerpt), jejichž databázové zpracování bude také přístupné na webu. Postupně budou přibývat další materiálové sbírky v elektronické podobě, např. rozsáhlý dialektologický archiv, soupisy pomístních jmen, sbírka osobních jmen aj.

Specifickou oblastí práce, která si vyžádá patřičnou pozornost, je **prezentace vědeckých výsledků pro širší veřejnost**. Vedle už existujících webových stránek ÚJČ plánujeme zprovoznění webového hnízda lexikograficko-terminologického oddělení, kde budou k dispozici sdružené aplikace s odstupňovanými uživatelskými právy (některé obsahové složky budou mít i nadále interní, ryze pracovní charakter). V současné době jsou již připraveny aplikace Databáze heslářů a Bibliografická databáze.

Databáze heslářů představuje souborný heslář vytvořený z heslářů PSJČ, SSJČ, SSČ, SN 1, SN 2 a FSČ. Pro potřeby analytické práce byly do databáze heslářů

Obr. 4. Databáze heslářů s vyhledaným řetězcem na konci slova

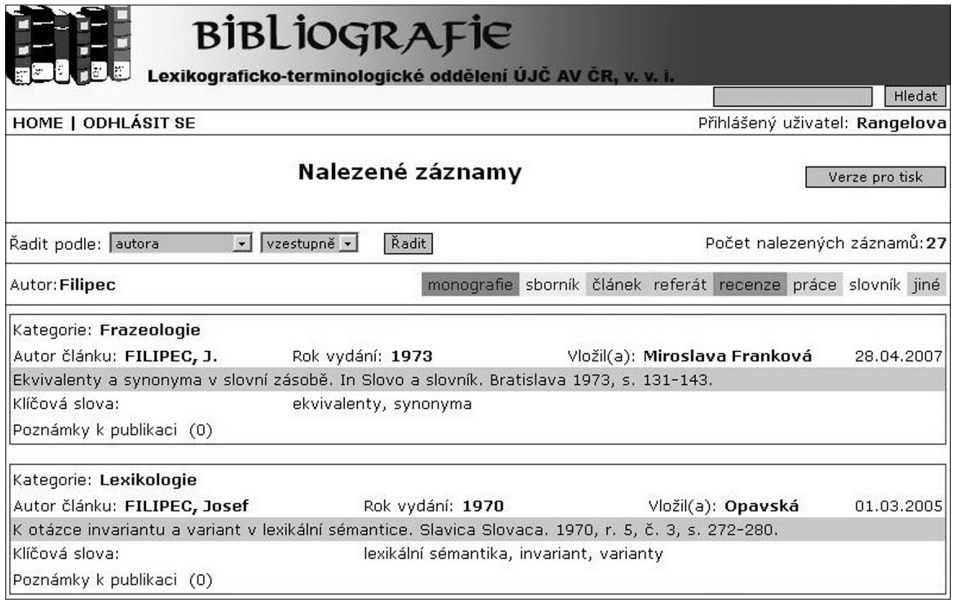
začleněny i některé nepublikované ucelené soubory slov, např. návrh hesláře lexikálního standardu[4]. V této databázi lze vyhledávat lemmata podle zadaného řetězce znaků s volbou jeho umístění (na začátku či na konci slova), viz obr. 4.

Za pomoci této aplikace je možné rychle získat informaci, ve kterém slovníku se hledaný výraz vyskytuje jako heslo. Zatím je možné kliknutím na zkratku názvu slovníku získat informaci o daném slovníkovém díle, v budoucnu se počítá s možností odskoku přímo do textu slovníku na hledané heslo. Databáze heslářů bude postupně doplňována dalšími hesláři a v budoucnu se stane východiskem pro tzv. lemmarium jako ucelený vyhledávací nástroj zahrnující hesláře materiálových sbírek, digitalizovaných slovníků a dalších, nejen lexikografických databází.

Bibliografická databáze LTO, která bude rovněž přístupná z budovaného webového hnízda, obsahuje strukturované bibliografické záznamy vztahující se k lexikografické práci. Je již funkční a plně k dispozici pro interní potřebu oddělení (viz obr. 5). Výhledově budou zpřístupněny texty příspěvků a studií členů našeho oddělení, související zejména s výzkumným záměrem, které byly publikovány v méně dostupných zdrojích (s uvedením bibliografického údaje).

Dále se počítá, že do té doby, než bude možné zveřejnit všechny materiálové sbírky ÚJČ, budou ve webovém hnízdě zveřejňovány ukázky materiálu – skeno-

4 Heslář lexikálního standardu vznikl v ÚJČ v 70. letech jako interní pracovní materiál pro úkol, který dále nepokračoval; je zachován jako rukopis bez uvedení autorů a bez datace.



Obr. 5. Bibliografická databáze s vyhledaným vzorkem záznamů

vaného a nově excerpovaného. Tyto ukázky budou doprovázeny informacemi o lexikálním archivu ÚJČ, příp. informacemi o jednotlivých lexikálních sbírkách. Zde bude možné rovněž umísťovat ukázky nového lexikografického popisu: v závěrečné fázi prací na výzkumném záměru (rok 2010) zde mohou být umístěna vybraná vzorová hesla z budované lexikální databáze, která budou připravena pro veřejnost a mohou plnit funkci dřívějších ukázkových sešitů s možností zpětné vazby (s využitím rubriky *Napište nám*).

Na závěr je třeba ještě podotknout, že tak náročný úkol, jakým je představený výzkumný záměr, předpokládá vytvoření odpovídajících technických a personálních podmínek pro dlouhodobou výzkumnou činnost, proto důležitými organizačními aspekty naší práce jsou budování výzkumného týmu (zapojení a vědecká průprava nových spolupracovníků) a optimální využití nových technologických možností (zabezpečení odpovídajících pracovních stanic a serveru, který by měl sloužit pouze potřebám výzkumného záměru). Díky novým technickým podmínkám je možné značně zefektivnit komunikaci a diskusi v rámci pracovního týmu: ve fázi vývoje je vnitřní komunikační prostředí s pracovním názvem Fórum, které bude sloužit i pro registrování a průběžné řešení dílčích problémů realizačního charakteru.

Výzkumný záměr *Vytvoření databáze lexikální zásoby českého jazyka počátku 21. století* ve své komplexnosti směřuje k vytvoření obsáhlého souboru lingvistických, zejména pak lexikálních dat, jejichž další využití bude mít zásadní vědecko-poznávací, dokumentační i národně a kulturně reprezentativní význam. V jeho

rámci se budují metodické, metodologické a technologické předpoklady moderní výzkumné práce v oblasti lexikologie a lexikografie, zaměřené na přípravu moderního výkladového slovníku češtiny. Plnění tohoto záměru přinese nejen specializované vědecké výsledky (rozsáhlé informace o české slovní zásobě), ale rovněž významně přispěje k lepší informovanosti naší i zahraniční veřejnosti o české lexikografické tradici i o současném výzkumu slovní zásoby.

Literatura

AKADEMICKÝ slovník cizích slov. Academia, Praha 1995. (ASCS)

ČESKÝ jazykový atlas 1. Academia, Praha 1992.

FREKVENČNÍ slovník češtiny. Nakladatelství Lidové noviny, Praha 2004. (FSČ)

MARTINCOVÁ, O. a kol.: Nová slova v češtině. Slovník neologizmů 1. Academia, Praha 1998. (SN 1)

MARTINCOVÁ, O. a kol.: Nová slova v češtině. Slovník neologizmů 2. Academia, Praha 2004. (SN 2)

MARTINCOVÁ, O. a kol.: Neologizmy v dnešní češtině. ÚJČ AV ČR, Praha 2005.

PALA, K. – HORÁK, A. – RAMBOUSEK, A. – RANGELOVA, A.: Nové nástroje pro českou lexikografii – DEB2. In: Gramatika a korpus. UJČ AV ČR, Praha 2007, s. 190–196.

PŘÍRUČNÍ slovník jazyka českého. Státní nakladatelství, Praha 1935–1957. (PSJČ)

RANGELOVA, A.: Izgraždane na baza za opisanie na posttotalitarnata leksika v češkija ezik. In: Ezikăt na totalitarnoto i posttotalitarnoto obštestvo, Prochazka i Kačarmazov, Sofia, 1996, s. 155–159.

SLOVNÍK spisovné češtiny pro školu a veřejnost. Academia, Praha 1978, 2. vyd. 1994, 3. vyd. 2003. (SSČ)

SLOVNÍK spisovné češtiny pro školu a veřejnost. Elektronická verze, LEDA spol. s r. o., Praha 2004. (SSČ-L)

SLOVNÍK spisovného jazyka českého. Nakladatelství Československé akademie věd, Praha 1960–1971, 2. vydání Academia, Praha 1989. (SSJČ)

SOCHOVÁ, Zd. – POŠTOLKOVÁ, B.: Co v slovnících nenajdete. Novinky v současné slovní zásobě. Portál, Praha 1994.

SVOZILOVÁ, N. – PROUZOVÁ, H. – JIRSOVÁ, A.: Slovesa pro praxi. Valenční slovník nejčastějších českých sloves. Academia, Praha 1997.

SVOZILOVÁ, N. – PROUZOVÁ, H. – JIRSOVÁ, A.: Slovník slovesných, substantivních a adjektivních vazeb a spojení. Academia, Praha 2005.

TĚŠITELOVÁ, M. – PETR, J. – KRÁLÍK, J.: Retrográdní slovník současné češtiny. Academia, Praha 1986.

VELKÝ slovník cizích slov. LEDA spol. s r. o., Praha 2005. (VSCS)

Textové zdroje

Český národní korpus – SYN2000. Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: <http://ucnk.ff.cuni.cz>.

Český národní korpus – SYN2005. Ústav Českého národního korpusu FF UK, Praha 2005. Dostupný z WWW: <http://ucnk.ff.cuni.cz>.

Český národní korpus – SYN2006PUB. Ústav Českého národního korpusu FF UK, Praha 2006. Dostupný z WWW: <http://ucnk.ff.cuni.cz>.

Textový archiv Newton Information Technology, s. r. o., www.newtonit.cz (Newton)