Práce s lexikálním materiálem a možnosti exemplifikace v lexikální databázi LEXIKON 21[1]

Zdeňka Opavská, Barbora Štěpánková
Ústav pro jazyk český AV ČR, v. v. i.

Working with Lexical Material and the Possibilities of the Exemplification in the Lexical Database LEXIKON 21
An integral component of the future treatment of entries in LEXIKON 21 is working with the lexical material and illustrative examples. This paper provides 1) a brief description of the lexicographic procedures for dealing with corpus material, and 2) a description of the tools for working with corpus evidence in the Lexikon 21: an independent program named PRAMAT (a 'lexicographer's "desktop" for working with the examples') and a special Exemplification tool within the PRALED database.

Neodmyslitelnou součástí budoucího zpracovávání hesel v novém výkladovém slovníku formou lexikální databáze nazvané *LEXIKON 21* (dále L 21)[2] bude – obdobně jako tomu bylo u výkladových slovníků PSJČ, SSJČ a SSČ – práce s lexikálním materiálem a exemplifikačními doklady. Základní postup při zpracování hesel bude týž: vychází se z lexikálního materiálu, z kontextů vztahujících se k dané lexikální jednotce, od nich se postupuje k významu/významům dané lexikální jednotky; jednotlivé významy a oblast jejich užití jsou pak dokládány v exemplifikační části heslové stati. Odlišná jsou však dnešní materiálová východiska i prostředky a nástroje, které lze při popisu české slovní zásoby použít. V mnoha ohledech se proto budou lišit i zásady zpracování pro L 21; to vyplývá ze samé podstaty tohoto nového lexikografického popisu.

Práce s lexikálním materiálem v L 21

L 21 se svými materiálovými zdroji liší od dosavadních výkladových slovníků. Na rozdíl od PSJČ, SSJČ a SSČ, které využívaly lístkový lexikální archiv, se pro

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

2 *LEXIKON 21* – lexikální databáze češtiny navrhovaná v lexikograficko-terminologickém oddělení ÚJČ AV ČR, v. v. i., pro zpracování budoucího výkladového slovníku; *L 21* je naše současné pracovní označení i pro tento budoucí výkladový slovník, který se v ÚJČ AV ČR začne zpracovávat až po roce 2011. Jednotlivé části koncepce postupně vznikají od roku 2005 v rámci současného výzkumného záměru *Vytvoření databáze lexikální zásoby českého jazyka počátku 21. století* (AV0Z90610521), viz též v tomto sborníku příspěvek J. Světlé *K návrhu, vývoji a funkcím lexikální databáze češtiny*.

L 21 stal základním materiálovým východiskem jazykový korpus psaných textů, a to *SYN2000*; v r. 2006 k němu přibyly další synchronní korpusy, a to *SYN2005* a *SYN2006PUB*[3]. Avšak obdobně jako se česká předpočítačová lexikografie opírala o propracované postupy pro práci s excerpovaným lexikálním materiálem, tak i pro zpracování hesel v rámci L 21 je nutné postupně vypracovávat a ověřovat lexikografické postupy pro práci s materiálem korpusovým. Na základě dosavadních zkušeností při zpracování zkušebního vzorku hesel byl takovýto postup navržen. Ve stručnosti lze říci, že jde o postup, v němž se kombinuje analýza konkordancí (v případě velkého množství konkordancí je používán náhodný vzorek o rozsahu 300 dokladů[4]) s analýzou kolokací získaných prostřednictvím nástroje Word Sketch, případně prostřednictvím korpusových statistických funkcí „Nejčetnější kolokace" a „Frekvenční distribuce". Pro analýzu konkordancí je zatím naším východiskem korpus *SYN2000*, v případě malé či nedostatečné doloženosti lexikální jednotky nebo lexikálního významu budou konkordance ze *SYN2000* doplněny o konkordance ze *SYN2005*, případně ze *SYN2006PUB*. Pokud jde o korpusové parametry, z údajů o zdroji jsme se rozhodli vybírat pro *SYN2000* tyto položky: doc.txtype, doc.temp, doc.opus; pro *SYN2005* a *SYN2006PUB* pak položky opus.txtype, opus.rokvyd, opus.id. Doporučená (nikoli však závazná) délka konkordancí je věta až několik vět[5]. Při zjišťování kolokací prostřednictvím Word Sketch lze využívat všechny tři zmíněné korpusy.

Uvedený pracovní postup bude během další lexikografické práce dále ověřován, doplňován a upravován.

Program *Pramat*[6]

Při práci s korpusem se ukázalo, že i když program Bonito umožňuje konkordance třídit a seskupovat do jednotlivých skupin na základě očíslování jednotlivých konkordancí, pro potřeby L 21 bude potřeba vyvinout specifičtější softwarový nástroj, který by sloužil jako lexikografova pracovní plocha pro práci s doklady. Na základě naší praxe byl navržen samostatný program Pramat; programátorem je Pavel Žikovský.

Pramat funguje nezávisle na Praledu a soubory v něm vytvořené se ukládají mimo prostředí Praledu. Bylo však navrženo, aby se vybrané doklady z Pramatu

3 Předpokládá se, že v budoucnosti budou tyto zdroje doplněny o další synchronní korpusy. Dalším zdrojem lexikálního materiálu bude také dílčí excerpce uskutečňovaná v lexikograficko-terminologickém oddělení ÚJČ.

4 Zde bude třeba ověřit, kdy a za jakých okolností bude nutné tento vzorek rozšířit o další doklady.

5 Tento rozsah byl doporučen vzhledem k přenosu vybraných konkordancí do Pramatu i vzhledem k tomu, že z Pramatu budou citátové (větné či vícevětné) doklady přenášeny do Exemplifikace v Praledu.

6 *Pramat* – samostatný program určený pro práci s doklady (zejména korpusovými), vyvíjený v lexikograficko-terminologickém oddělení ÚJČ AV ČR, v. v. i.

přenášely (kopírovaly) do nástroje Exemplifikace v programu Praled v předem stanoveném formátu. Na vývoji Pramatu a jeho úpravách se dále pracuje.

Tabulková část Pramatu je primárně určena pro práci s doklady (maximální počet dokladů v jednom souboru je 1000) a umožňuje vkládání dokladů, jejich komentování, třídění, mazání, kopírování a přenášení do nástroje Exemplifikace v Praledu. Doklady v programu Pramat je možné setřídit podle identifikačního čísla dokladu (ID), podle polí určených pro okomentování dokladu (Tag1, Tag2, Tag3) a polí obsahujících informace o zdroji dokladu (Zdroj1, Zdroj2). Funkce Vložit z Bonita umožňuje automaticky vložit a rozdělit vybrané korpusové doklady do polí určených pro text dokladu a do polí určených pro údaje o zdroji dokladu.

Do poznámkové části Pramatu si může zpracovatel poznamenat cokoli, co pro svou práci potřebuje, např. pracovní výklady významu, pracovní poznámky atd.

V návrhu je možnost kopírování a ukládání vybraných dokladů do nástroje Exemplifikace v Praledu v předem daném formátu a rovněž se počítá s tím, že se po kliknutí na stručné údaje o zdroji zobrazí jejich plné znění. Toto rozkódování zdrojových údajů by pak mělo fungovat jak v Pramatu, tak i v Exemplifikaci v rámci Praledu.

Exemplifikace v *Praledu*[7]

Exemplifikace ve výkladových slovnících slouží (1) k ilustraci užívání lexému v kontextu, (2) k verifikaci jeho významu a (3) k informování o jeho lexikální i gramatické spojitelnosti.[8] Požadavky na exemplifikaci lze nalézt např. u F. Čermáka: „Exemplifikace lemmatu musí být především jak typická, tj. ilustrující typický (a tedy nikoliv často idiosynkratický autorský) úzus, tak konkretizační vůči veškeré informaci většiny hesla, zvláště však vůči informaci naznačené funkcí, valencí, kontextem a sémantikou…“[9].

Výkladové slovníky PSJČ, SSJČ a SSČ se materiálově opíraly o dlouhodobě budovaný a doplňovaný lístkový lexikální archiv Ústavu pro jazyk český. Způsob využití tohoto lexikálního materiálu při zpracování slovníkových hesel a výběr dokladů do exemplifikace byl dán – stejně jako jiné slovníkové aspekty – zaměřením, rozsahem a koncepcí daného slovníku[10]. Tak byl PSJČ koncipován jako slovník citátový (jen zřídka byly využity doklady „necitátové“), v rozsahově středním SSJČ

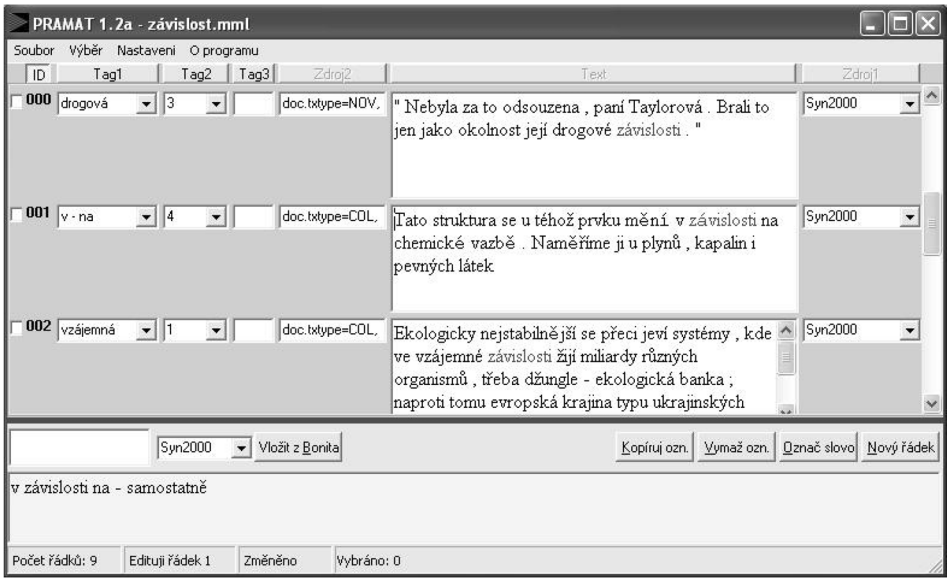
---

7 *Praled (Pražská lexikální databáze)* – program vyvíjený ve spolupráci lexikograficko-terminologického oddělení ÚJČ AV ČR, v. v. i., a Centra zpracování přirozeného jazyka FI MU v Brně.

8 Srov. např. Filipec, Manuál lexikografie 1995, s. 37–40, Čermák, tamtéž, s. 107–108.

9 Manuál lexikografie 1995, s. 107.

10 O historii a koncepci českých výkladových slovníků souhrnně viz Filipec 1975, Hladká 2005. Stručně o koncepci daných slovníků viz též stati o zpracování slovníku v 1. díle PSJČ, v 1. díle SSJČ a v SSČ.



Obr. 1. Nástroj pro práci s doklady – program PRAMAT

byly – vzhledem k jeho informativnímu zřeteli – kombinovány necitátové doklady s citátovými (citátové doklady především dokumentovaly slova nebo významy z periferie slovní zásoby) a v jednosvazkovém SSČ byly použity – až na několik výjimek – pouze necitátové doklady (důraz byl kladen na ilustraci typičnosti užití).[11] Rozsah tištěného slovníku byl vždy důležitým limitujícím faktorem jak pro samotný způsob exemplifikace (viz např. zkracování heslového slova v exemplifikaci SSJČ i SSČ), tak pro množství uváděných dokladů (srov. poměrně bohatou exemplifikaci v SSJČ s minimalizovanou exemplifikací v SSČ, ve které je „uvádění příkladových spojení (…) omezeno na nejnutnější míru“[12]).

Při zpracování exemplifikace[13] v L 21 navrhujeme využívat oba hlavní typy dokladů, tj. jak doklady citátové (=větné), tak doklady necitátové (=upravené, příkladové). Výhodou databáze je možnost vložení velkého množství dokladů, počítá se proto s tím, že ve slovníkovém hesle L 21 budou použity v relevantních případech oba typy exemplifikace. Ukazuje se totiž, že například pro exemplifikaci synsémantik, citoslovcí či některých typů příslovcí je vzhledem k jejich funkci v textu (a např. i kvůli homonymii) nutné uvádět především citátovou exemplifikaci.

11 Problematice zpracování exemplifikace v PSJČ, SSJČ a SSČ byl věnován podrobný rukopisný příspěvek J. Machače (nedatováno). O podobě exemplifikace viz též zásady zpracování v jednotlivých slovnících.

12 SSČ, Zásady zpracování slovníku, 1978, s. 795.

13 Pojetí exemplifikace pro L 21 vzniká jako společné dílo kolektivu lexikograficko-terminologického oddělení ÚJČ AV ČR, v. v. i.

E2 k V1 **(klikněte pro rozbalení/sbalení)**

syn2000|doc.txtype=NOV,doc.temp=1995,doc.opus=brizy Dovedu být šetrná . Jedny šaty a < boty> mi vydrží rok.

syn2000|doc.txtype=NOV,doc.temp=1996,doc.opus=ledsyn Oblékla si elegantní šaty a kabát z velmi jemné kůže a obula si < boty> na podpatku .

syn2000|doc.txtype=PUB,doc.temp=1999,doc.opus=mf990820 Do lesa znalci doporučují nechodit naboso a v krátkých kalhotech. " Vhodné jsou pevné < boty> a dlouhé kalhoty," říká Vitáček.

syn2000|doc.txtype=COL,doc.temp=1994,doc.opus=pi310 Heřman se zul , protože ho < tlačily boty>...

Obr. 2. Citátové doklady v exemplifikačním bloku

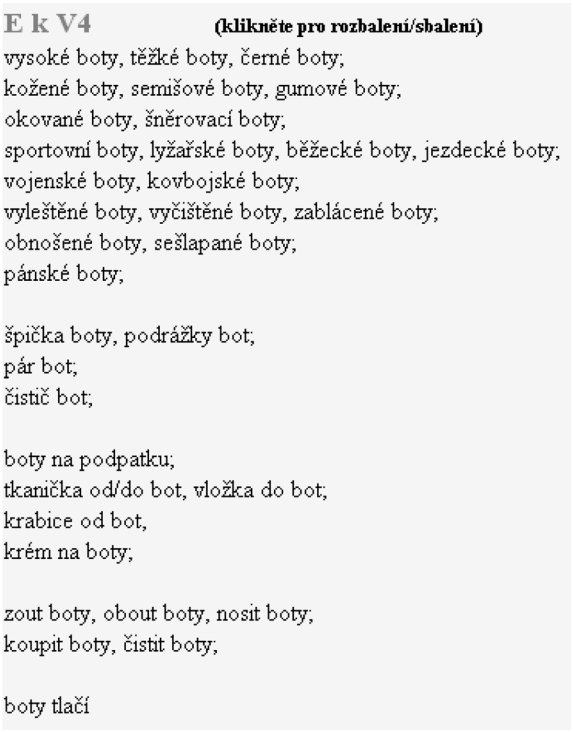
Citátové doklady budou mít v L 21 podobu větných nebo vícevětných kontextů s uvedením zdroje. V L 21 se na rozdíl např. od PSJČ nebudou objevovat jen citáty z beletrie, ale vhodné doklady mohou být v korpusech vybírány z různých žánrů. V citátových dokladech se tak projeví snaha podchytit současný úzus (alespoň co se týče psaných veřejných textů). Jako informace o zdroji dokladu se v exemplifikaci zobrazí název korpusu, typ textu, rok vydání a zkrácený název zdroje.

Označení **necitátové doklady** v podstatě odpovídá exemplifikaci příkladovými spojeními ve formě syntagmat. Jde především o kolokace[14] (slovní spojení). Touto formou budou zachycovány zejména typické a běžné kolokace; jejich řazení bude odpovídat formálně-sémantickému modelu, konkrétní kolokace pak vybere autor na základě statistických a kolokačních nástrojů (frekvence, soudržnost kolokací, word sketches apod.); samozřejmě i zde se autor a následně redaktoři musí opírat též o své jazykové povědomí.

Například pro řazení typických a běžných kolokací u substantiv byl pro L 21 navržen tento základní model: 1. adjektivum+substantivum, 2. substantivum+substantivum, 3. substantivum+předložkový pád, 4. verbum+substantivum, 5. substantivum+verbum, 6. koordinační spojení[15]. Model slouží pouze jako podklad pro autory, nepředpokládá se totiž, že u všech substantiv se budou vyskytovat všechny položky modelu. Je-li to možné, budou lexémy v exemplifikaci uváděny v reprezentativním tvaru, zároveň se však bude přihlížet ke skutečnému užívání, např. objevují-li se některé kolokace výhradně v množném čísle, uvedou se v této podobě i zde.

[14] Kolokaci chápeme jako „syntagma jaz. prvků lexikální povahy“ (F. Čermák, Z. Hladká v ESČ, s. 218), nikoli jen jako „typické kolokace“.

[15] Předpokládáme, že koordinační spojení se budou v L 21 objevovat i v podobě citátových dokladů, jejich řazení mezi nevětné kolokace je spíše zpracovatelskou pomůckou.



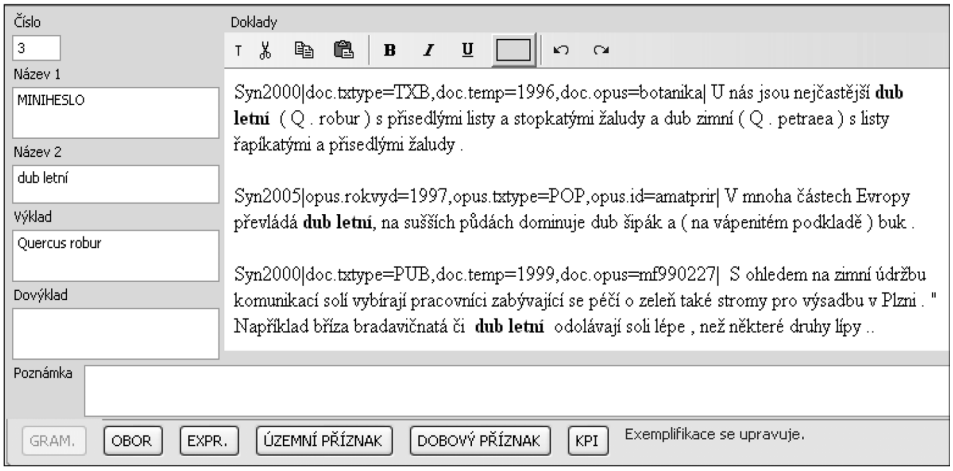
Obr. 3. Řazení necitátových dokladů v exemplifikačním bloku

Otevřenou otázkou zůstává užívání tzv. delších upravených dokladů – tyto větné struktury zachycující nějaký důležitý rys lexikální jednotky by bylo možno užívat např. k doložení valence slovesa. Šlo by tedy o doklady, jejichž primárním účelem by nebylo zachycení skutečného textového užití, ale zdůraznění určitého jevu.

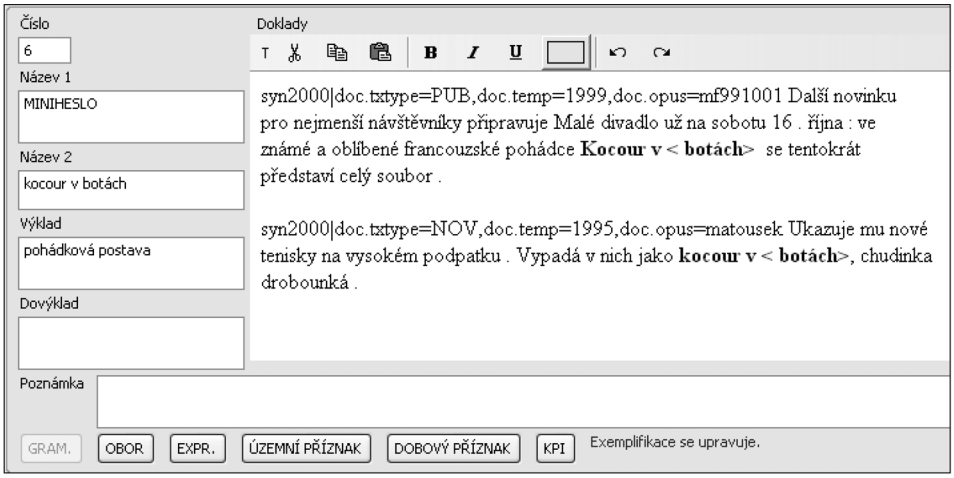
Nástroj Exemplifikace je v Praledu strukturován do tzv. exemplifikačních bloků s jednotnou strukturou v každém z nich: Číslo, Název 1 (pro označení typu exemplifikace), Název 2 (pro uvedení konkrétní kolokace), Výklad (pro výklad významu u tzv. minihesla), Dovýklad (pro doplňující výklad k dokladu), Doklady, nabídka kvalifikátorů (gramatika, obor, územní příznak atd.) a Poznámka.

V každém bloku budou zaznamenávány doklady obdobného charakteru dokládající jednotlivé typy užití. Pokud jde o pořadí bloků, budeme zřejmě postupovat od nepříznakových dokladů k příznakovým. Na prvním místě budou v samostatných blocích uváděny doklady, jejichž hlavní funkcí je ukázat na typické a běžné užití slova, tedy ilustrovat spojitelnost, a dále doložit výklad významu. V následujících samostatných blocích budou uváděny doklady, které vyžadují specifické doplnění: doklady s příznakovým užitím, doklady s přeneseným užitím, doklady zaznamenávající specifický gramatický či sémantický rys.

Uvažujeme o tom, že zvláštním typem exemplifikace se může stát tzv. *miniheslo*. Tento exemplifikační blok má dvě různé funkce. Jednak může sloužit k za-



Obr. 4. Zachycení víceslovného termínu v minihesle



Obr. 5. Využití tzv. registrační funkce minihesla

chycení určitých typů víceslovných termínů, např. těch, které v hierarchickém řazení daného oboru specifikují konkrétní druh (např. u lexému *dub* v terminologickém významu ‚*botanický rod stromů Quercus*' jsou minihesla pro *dub letní* a *dub zimní*). Jednak lze hovořit o tzv. registrační funkci. V minihesle se mohou zachycovat i kolokace, které jsou na přechodu k samostatnému heslu (tj. k frazému či jiné víceslovné lexikální jednotce). Díky možnostem formuláře lze miniheslo opatřit výkladem, dostatečným množstvím příkladů, gramatickou informací i kvalifikátory; je tedy snadno transformovatelné do podoby samostatného hesla.

Zaměření L 21 na korpusové zdroje s sebou nese, vedle možnosti práce s daleko větším množstvím různorodých dokladů a tím i možností získat rozsáhlejší oporu v úzu, i větší požadavky na nástroje, které tuto práci umožní i usnadní. Mezi tyto nástroje patří i program Pramat, který je vytvářen přímo pro specifické potřeby zpracovatelů budoucího výkladového slovníku, zejména co se týče třídění vybraných kontextů a ukládání dokladů do exemplifikační části hesla v L 21.

Nástroj Exemplifikace je pro slovníkovou databázi L 21 navržen tak, že umožňuje nejen dokládání jednotlivých významů a významových odstínů pomocí dostatečného množství citátových i necitátových dokladů, ale díky flexibilní databázové struktuře se také snadno přizpůsobí i dalším zpracovatelským požadavkům, např. při zpracovávání víceslovných lexikálních jednotek apod. Do budoucna se samozřejmě počítá s dalším vývojem obou nástrojů.

Literatura

ENCYKLOPEDICKÝ slovník češtiny (ed. P. Karlík, M. Nekula, J. Pleskalová). Praha, Nakladatelství Lidové noviny 2002. (ESČ)

FILIPEC, J.: Cesta k českému jednosvazkovému slovníku. Naše řeč 58, 1975, č. 5, s. 225–233.

HLADKÁ, Z.: České slovníkářství na cestě k jednojazyčnému výkladovému slovníku. Naše řeč 88, 2005, č. 3, s. 140–159.

MACHAČ, J.: Slovní spojení (v jednojazyčném výkladovém slovníku). (rkp., nedatováno, 10 stran)

MANUÁL lexikografie (ed. F. Čermák, R. Blatná). Jinočany, H&H 1995.

PŘEDMLUVA. In: Příruční slovník jazyka českého. A-J. Praha, Státní nakladatelství 1935–1937, s. VII–XI.

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

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

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

VÝKLAD o uspořádání slovníku. In: Slovník spisovného jazyka českého. A-M. Praha, Nakladatelství Československé akademie věd 1960, s. VII–XVIII.

ZÁSADY zpracování slovníku. In: Slovník spisovné češtiny pro školu a veřejnost. Praha, Academia 1978, s. 779–799.

ZÁSADY zpracování slovníku. In: Slovník spisovné češtiny pro školu a veřejnost. Praha, Academia 2003, s. 641–646.

Korpusy a programy:

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>.

Word Sketch Engine, dostupný z WWW: <http://ucnk.ff.cuni.cz/corpora>.