Stručni članak
UDK 81'374:004.65:004.738.5

Primljen 1. 6. 2001.
Prihvaćen 14. 6. 2001.

LEKSIČKE BAZE PODATAKA – IZVORI S INTERNETA

Boris Pritchard *
Sveučilište u Rijeci, Odjel za pomorstvo i Visoka pomorska škola

U uvodu u leksičke baze podataka daje se definicija i njihova klasifikacija uz prikaz nekih teoretskih pitanja. Prikazane su odabrane baze podataka i navedeni njihovi izvori i adrese na internetu.

Ključne riječi: Internet, leksičke baze podataka

Riječi su ključevi kojima se potiče kreativno i analitičko mišljenje o odnosu sadržaja (stvari) i jezične forme (riječi). Put od stvari do riječi nije izravan nego se ostvaruje putem konceptualizacije, stvaranja pojma o svijetu oko nas i o nama samima. Leksičke baze podataka sadrže pojmove kojima se stvara naša mentalna slika o svijetu, odnosno koncepualizira se izvanlingvistička stvarnost. To je znanje formalizirano u obliku:

- logičkih podataka, odnosno logičkih informacija o pojmovima (skup deduktivnih pravila koja se primijenjuju na logičke oblike koje taj pojam tvori)
- informacija o sadržaju pojma (enciklopedijskih podataka)
- leksičkih podataka
- metalingvističkog nazivlja pojmova (služe za kategorizaciju pojmova i prikazivanje njihovih međusobnih odnosa) te
- njihovih leksikaliziranih oblika i leksičkih podataka u prirodnom jeziku, odnosno informacija o realizaciji nekog pojma u prirodnom jeziku.

Valja ih razlikovati od rječničkih baza podataka (*dictionary database*), koje sadrže skup reprezentativnih tekstova za leksikografsku obradu leksičkih jedinica prirodnih jezika. Takve su baze npr. *Bank of English (Collins-COBUILD http://titania.cobuild.collins.co.uk*, Sveučilišta u Birminghamu), BNC (British National Corpus *http://info.ox.ac.uk/bnc*), HNK (Hrvatski nacionalni korpus; *www.hnk.ffzg.hr*) itd.

Broj pojmovnih natuknica u leksičkim bazama podataka relativno je nizak (danas najviše sedam do osam tisuća pojmova), ali su vrlo opsežne informacije o sadržaju i semantičkim odnosima među pojmovima, pa stoga vrlo intenzivno koriste uputnice prema drugim pojmovima, kategorijama, sinonimima, nadređenim i podređenim pojmovima. Nasuprot njima tekstualne rječničke baze u pravilu sadrže više milijuna riječi (pojavnica), a njihov broj ovisi o namjeni baze podataka, stupnju reprezentativnosti i kapacitetima elektroničkog pohranjivanja, odnosno o operacijskim mogućnostima suvremenih računalnih programa i aplikacija.

Tezaurus je najstariji pojmovni oblik u leksikografiji, a odnosi se na posebnu vrstu rječnika ustrojenoga prema sinonimijskim odnosima, hijerarhijskim vezama i asocijativnim odnosima unutar nekog tematskog područja, odnosno to je sustavni popis sinonima i njihovih skupova te s njima povezanih pojmova. Isprva (16. st.) riječ tezaurus označavala je 'spremište, riznicu' a zatim 'rječnik' ili 'enciklopediju', da

* Boris Pritchard, Sveučilište u Rijeci, Odjel za pomorstvo i Visoka pomorska škola, Studentska 2, 51000 Rijeka; telefon: (051) 338-411; telefax: (051) 336-755; e-mail: bopri@pfri.hr

bi se od 1852. godine, kada je bio objavljen Rogetov *Thesaurus of English Words and Phrases*, rabila u današnjem značenju 'pojmovnika' u užem, specifičnom smislu rječnika sinonima i širem, generičkom smislu, pojmovnog rječnika utemeljenog na sinonimiji i hijerarhijskim odnosima. Pojmovi i pojmovne informacije nalaze se, međutim, gotovo u svim vrstama rječnika i pojmovnika, posebno u definicijama u jednojezičnim rječnicima, pojmovnim i kategorijanim odrednicama u jednojezičnim i dvojezičnim rječnicima, a čine osnovu organizacije tematskih rječnika.

U ovom se članku daje uvod u teoretsku podlogu leksičkih baza podataka i njihovo razvrstavanje, a na kraju se navode izvori za pronalaženje i pretraživanje (*on-line*) leksičkih baza podataka na Internetu, odnosno odgovarajuće www-stranice leksičkih baza podataka, uz kraća objašnjenja. Ovdje se pod pojmom leksička baza podataka (*lexical database*)[1] podrazumijeva prvenstveno njihov računalni oblik (računalni pojmovnik). Računalni pojmovnici mogu biti opće namjene i služe za strukturiranje općejezičnih leksičkih baza podataka (npr. WordNet 1.6, EuroWordNet, EAGLES), popratni (uz jednojezične rječnike: CEDT, WNW, RHD) te predmetni, strukovni ili 'kontrolirani' tezaurusi (UNESCO Thesaurus, RCHME Thesaurus itd.).

Današnjem razvitku leksičkih baza podataka, tj. kategoriziranih popisa pojmova-izraza radi pronalaženja informacija računalnim putem ('a categorized index of terms for use in information retrieval, as from a computer'), koje su, sastavljene, strukturirane i vođene posebno izrađenim pojmovnim sustavima, pridonijele su, naročito nakon sedamdesetih godina, generativna semantika (Fillmore, Jackendoff, Katz), Chomskyjeve teorije Government & Binding, FG i LFG (Halliday), leksičke semantike (Cruse, Jackendoff, Talmy, Dik, Fillmore, Lakoff, Mel'cuk, Pustejovsky, Wierzbycka itd.,), psiholingvistika i kognitivna znanost, logika te informatičke znanosti, posebice umjetna inteligencija (AI) i sistemske znanosti (knjižničarski sustavi, UDK itd.). Dok je ranije rasprava o sistematizaciji pojmova o ukupnom ljudskom znanju uglavnom bila u domeni filozofije, danas je ona pretežito predmet sistemskih znanosti, organizacije velikih svjetskih informatičkih sustava (Information Systems), npr. bibliotečnih i klasifikacisjkih sustava, te psiholingvistike. O opisu i kategorizaciji pojmova vidi M. Hearst (1998) *Categories and Attributes* (*www.sims.berkeley.edu*) a o leksičkoj semantici: Cruse 1986 te SIGLEX Resources, ACL SIGLEX, Special Interest Group on the Lexicon of the Association for Computational Linguistics (*http://www.clres.com/siglex.html*).

Svaka računalna leksička baza podataka unutar pojedine natuknice (pojma,) sadrži ove konceptualne informacije (cf. Sperber and Wilson 1986:86):

(a) logičke informacije o pojmu, tj. skup deduktivnih pravila koja se odnose na logičke oblike kojih je sastavnica taj pojam po sebi,

(b) enciklopedijske informacije o opsegu i značenju pojma, tj. o stvarima, događajima i svojstvima koji obilježavaju taj pojam,

(c) leksičke informacije o leksikaliziranom obliku tog pojma u određenom prirodnom jeziku, tj. riječ ili izraz nekog jezika koja izražava taj pojam.

Po načelu logičko-filozofske kategorizacije spoznaje i znanja o svijetu svaki pojam pripada i sadržan je u višem, općenitijem pojmu (*genus proximum*), a istodobno ima i konkretnije osobine kojima se razlikuje od drugih pojmova (*differentia specifica*). Tako će pojam u leksičkoj bazi podataka, osim spomenutih informacija sadržati i informacije lingvističke prirode (semantičke i sintaktičke). Stoga u ljudskom komuniciranju tako nastali pojmovi, a primjereno tome i leksičke (pojmovne) baze podataka, imaju osnovnu funkciju da nam, potpomognute suvremenim dostignućima kognitivne znanosti (*cognitive science*) i umjetne inteligencije (*AI, artificial inteligence*), pružaju izvjesnu stabilnost u interakciji s okolinom. Također, omogućuju nam da se izdignemo iznad specifičnih pojedinosti naše okoline

te da stvari ili pojave koje imaju zajednička svojstva smatramo jedinicama koje pripadaju određenoj vrsti (Ellis, H.C. & Hunt, R.R., 1972/ 1993: 204 – 222).

Pojmovi su u leksičkim bazama strukturirani, tj. razvrstani su i organizirani, po logičkim i semantičko-sinatakičkim načelima. Kategorije su osnovni logički, kognitivni i psiholingvistički oblik strukturnog organiziranja pojmova. Kategorijalni pojmovi su osnovni (najopćenitiji) pojmovi koji nisu izvedeni iz drugih pojmova nego su zamisli osnovnih vrsta predmeta i iz njih se izvode ostali pojmovi (*materija, kretanje, prostor, vrijeme, stvar, uzrok, identičnost, proturječnost, kvaliteta, količina, procesi, sadržaj, oblik, odnos itd.*).

Semantički su odnosi temelj svake organizacije pojmova unutar leksičke baze podataka. Većina leksičkih baza podataka polazi od sinonimije kao temeljnog semantičkog odnosa kako bi se došlo do osnovne organizacije pojmovlja utemeljene na različitim vrstama hijerarhijskih, odnosno inkluzijskih odnosa: metonimije, antonimije, taksonomije, uzrok-posljedica, vršitelj – radnja itd. Osnovno je načelo funkcioniranja leksičke baze podataka da zadanim pojmom ili nekom njegovom kognitivnom ili semantičkom sastavnicom, putem indeksacije odnosno kazala leksičkih i semantičkih odnosa (*word relations index*), možemo doći do drugih pojmova iz istog semantičkog pa i asocijativnog područja te tako stvoriti cjelinu (predodžbu, sliku). To uključuje i složenije odnose, npr. konotacije, asocijacije indirektne leksičke veze, pa i slučajne asocijacije. Upravo indeksiranjem i leksičko – konceptualnom analizom leksičke baze podataka omogućuju vrlo brze operacije velikom količinom informacija, tj. značenjsko-funkcionalne operacije s leksičkim informacijama. O teoretskim pristupima projektiranju leksičkih baza podataka vidi Miller et al. (1993) na www adresi *www.cogsci.princeton.edu~wn/Publications*).

Danas se već intenzivno istražuju mogućnosti uključivanja složenijih sintaktičkih, semantičkih te sociolingvističkih i pragmalingvističkih informacija u leksičke baze podataka.

Budući da danas u svijetu postoje brojni projekti izrade i održavanja leksičkih baza podataka, nameće se potreba ubrznog donošenja leksičkih semantičkih standarda za jezični inženjering (*human language technology, HLT*), kao što su konceptualna i semantička obrada leksika, vrednovanje, kodiranje / mark-up, govorni jezik itd. To je zadaća posebnoga projekta: *The Expert Advisory Group on Language Engineering Standards* (EAGLES), DG XIII (Eur. Commis, *Linguistic Research and Engineering*). Rad je na projektu započet 1993. godine, a koordinator / nositelj ftp-servera je Consorzio Pisa Ricerche i može se naći na www adresama:

(www.linglink.lu/le/projects/EAGLES/ index. html) i (www.ilc.pi.cnr.it/EAGLES/ home.html)

Dosadašnji rezultati rada na najtežem pitanju: standardizaciji klasifikacije pojmova i njezinom usklađivanju na različitim područjima ljudske djelatnosti i znanosti: npr. u suvremenoj psiholingvistici, kognitivnoj znanosti te u sistemskim znanostima (*AI, information retrieval, Library Categorization*) dostupni su na stranicama EAGLES-a. Neuhvatljivost pojma „značenje" ostaje, međutim, i nadalje osnovna teškoća i boljka svakog pojmovnog ususta-vljivanja leksika.[2]

Leksičke baze podataka ili računalni leksikoni (*computational lexicons*) danas pružaju informacije kao što su:

- morfosintaktičke opise i podatke
- podatke o sintaktičkoj i semantičkoj subkategorizaciji
- opće opise leksičke naravi u korpusima i rječnicima[3]

Leksičke baze podataka namijenjene su za obradu tekstova i leksičkih podataka, sustavno klasificiranje, organiziranje, pretraživanje i vođenje informatičkih sustava, za generiranje prirodnih jezika (*NLP – Natural language Processing*) i njihovih rječnika, strojno prevođenje, traženje informacija (*Information Extraction*) itd.

U najšire shvaćenoj podjeli mogu se svrstati u leksičke baze podataka (LDB):

(a) opće namjene (za strukturiranje općejezičnih leksičkih baza podataka: WordNet 1.6, EuroWordNet, WordSmith),

(b) popratne konceptualne sustave (uz definicije i za generiranje definicija u jednojezičnim rječnicima: CEDT, WNW, RHD) te

(c) predmetne, strukovne ili 'kontrolirane' tezauruse (UNESCO Thesaurus, RCHME Thesaurus itd.).

One su, nadalje, namijenjene za suvremena teoretska jezična istraživanja, ali i za primijenjena istraživanja na području jezičnog inženjeringa (*human language engineering, HLT*) te za analize leksika i teksta.

Sveobuhvatan pregled, pa čak i puko navođenje naziva danas postojećih leksičkih baza podataka daleko nadilazi predviđeni prostor, pa se ovdje, prema vlastitu izboru autora i prema vlastitoj klasifikaciji, uz kraći komentar, navode sljedeće vrste leksičkih baza podataka:

1. KLASIČNI RJEČNICI KAO LEKSIČKE BAZE

1.1. The Longman Lexicon of Contemporary English:

http://www.ilc.pi.cnr.it/EAGLES96/rep2/node18.html

http://www.books.lt/CAMBRIDG/iner.htm

Računalni rječnici (*Machine-Readable Dictionaries*) *The Longman Dictionary of Contemporary English* te *The Longman Dictionary and Thesaurus* često se koriste kao izvori za kreiranje leksičkih baza podataka (*The Longman Dictionary and the Longman Lexicon of Contemporary English*) za prirodne jezike. Iako im je organizacija i struktura tradicionalna, neka njihova obilježja, posebno definicije, vrlo su pogodna za primjenu u izradi pojmovno organiziranih leksičkih baza podataka.

2. DVOJEZIČNI ELEKTRONIČKI RJEČNICI (BILINGUAL DICTIONARIES)

2.1. The bilingual Oxford Hachette French dictionary

www.oup.co.uk/isbn/0-19-268307-1

2.2. Van Dale Bilingual Dutch-English

www.tue.nl/bib/cdvdneen,leftmenu,content.html

Oba navedena elektronička rječnika vrijedna su za računalne lingviste, jer su mnogi leksički podaci u tim rječnicima obilježeni univerzalnim sustavom obilježavanja riječi (sgml). Obilježene su vrste riječi, homografi, različita značenja, izgovor, upotreba, idiomatski izrazi, područje (*domain*), kolokacije subjekt-objekt, prijedložni izrazi itd., što je iznimno važno za računalne operacije u sklopu leksičkih baza podataka. Sadrže i elektronički obilježene skraćenice, akronime, afikse, vlastita imena.

3. NAMJENSKE LEKSIČKE BAZE

3.1. GLDB – The Göteborg Lexical DataBase

www.ilc.pi.cnr.it/EAGLES96/rep2/node19.html

Ovu leksičku bazu podataka vodi od 1997. Odjel za švedski jezik Sveučilišta u Göteborgu. U osnovi je lingvističkog modela 'lema-leksem', gdje 'lema' predstavlja kanonički oblik riječi (i svih njenih oblika te podataka o njima) dok 'leksem' daje semantički sadržaj i podjelu na značenja. Leksemi se dijele u dvije kategorije: obvezna jezgrena značenja i neobvezna podznačenja. Obje kategorije sadrže varijable: definicije, formalne (gramatičke) komentare, reference te morfo-sintaktičke primjere. GLDB obuhvaća čitav jezik i iz te su se baze razvila dva najveća švedska jednojezična rječnika. Postoji i verzija na CD-romu. Osnovu semantičkih in-

formacija čine definicije aristotelovskog tipa koje uključuju *genus proximum* i *differentia specifica*.

4. LEKSIČKE MREŽE (wordnets)

4.1. The Princeton WordNet 1.5 i 1.6, 1.7

www.cogsci.princeton.edu/~wn

www-mitpress.mit.edu

WordNet je on-line leksička baza podataka nastala razvitkom suvremene teorija psiholingvistike i kognitivne znanosti i teorije o čovjekoj memoriji. Vrste riječi engleskog jezika ustrojene su u sinonimne skupove (*synsets*), od kojih svaka predstavlja jedan leksički pojam. Sinonimni skupovi međusobno su povezani i upućuju jedni na druge. Među sinonimnim skupovima vladaju sljedeći odnosi: sinonimija, hiperernimija, hiponimija, holonimija, antonimija itd. Uz synsete dane su i sintaktičke informacije. Bazu WordNet razvio je Cognitive Science Laboratory na sveučilištu Princeton pod vodstvom profesora Georgea A. Millera, glavnog istraživača (*Principal Investigator*). Baza se može pretraživati on-line ili na CD-romu. Ova je baza danas najvažniji izvor istraživačima u računalnoj lingvistici, analizi teksta i drugim srodnim područjima. Namijenjena je identifikaciji značenja riječi te traženju i dobivanju leksičkih informacija.

4.2. EuroWordNet

www.hum.uva.nl/~ewn/

EuroWordNet je višejezična leksička baza podataka za nekoliko europskih jezika (češki, estonski, francuski, holandski, njemački, španjolski i talijanski). Polazna mu je baza Wordnet, američka leksička baza za engleksi jezik, te koristi njegove sinonimske skupove. Svaka je mreža jedinstven sustav leksikalizacije unutar određenog jezika. Sve su baze povezane na WordNetov program Inter-Lingual Index te je omogućeno povezivanje među pojedinim jezicima. Taj indeks također omogućava pristup najvišoj ontologiji (pojmovima najvišeg reda) sa 63 semantička obilježja. Ontologija je zajednička za sve jezike, a specifičnosti pojedinih jezika zadržane se u njihovim vlastitim leksičkim mrežama. Baze se koriste za dobivanje jednojezičnih i dvojezičnih informacija. Za sada ova leksička baza nije dostupna besplatno na Internetu, pa se zainteresirani upućuju na (skupu) pretplatu za periodično korištenje.

5. LEKSIČKE BAZE PODATAKA ZA STROJNO PREVOĐENJE (Lexicons for Machine-Translation)

5.1. Eurotra Lexical Resources

http://www.ccl.kuleuven.ac.be

5.2. METAL Lexical Resources

www.ccl.kuleuven.ac.be/about/METAL.html

5.3. Logos Lexical Resources

http://www.ilc.pi.cnr.it/EAGLES96/rep2/node25.html

http://lrc.csis.ul.ie/LRC/presentations/...gos/tsld003.htm

5.4. Systran Lexical Resources

http://www.systransoft.com

Systran je sofisticirani program strojnog prevođenja u kojemu je proces prevođenja utemeljen na rekurzivnom pretraživanju (skeniranju) riječi u rečenici kako bi se došlo do prihvatljivih odnosa među leksičkim jedinicama rečenice. Sustav koristi rječnike da bi definirao leksičke jedinice analizirajući morfeme i kombinirajući njihov morfološki, sintaktički, semantički i propozicijski sadržaj. Riječi se semantički enkodiraju (i obilježavaju) radi određivanja sintaktičkih funkcija (*parsing*), npr:

ABSUB = Verb normally takes an abstract subject

ANSUB = Verb normally takes an animate subject

COSUB = Verb normally takes a concrete subject

HUSUB = Verb normally takes a human subject

Imenice se također semantički i konceptualno obilježavaju (markiraju) i razvrstavaju npr.: CON = Conrete, CT = Countable, MS = Mass, HU = Human, TP = Time Period itd. Slično je i s prilozima (TI = Time, PL = Place, MA = Manner, FREQ = Frequency, MODA = Modality, DIR = Direction, FUT = Future time). Semantičke informacije čine i dio složenijih informacija (semantički primitivi i terminološki kodovi, informacije o predmetu, području ljudskog znanja te vrsta teksta) itd.

6. ONTOLOGIJE VIŠEG REDA (higher level ontologies)

6.1. Cycorp

(izg. /'salko:/), *http://www.cyc. com/* jest baza podataka koja daje uvod u ontološki sustav programa CYC®. Sadrži oko 3000 pojmova iz vlastite baze znanja o svijetu, razvrstano u 43 glavne skupine (*topical groups*). Uz klasifikaciju pojmova daju se i upute o sintaksi programa (*The Syntax of CycL*) te glosar osnovnih termina. Oko milijun ručno unesenih tvrdnji (*assertions*) obuhvaća velik dio znanja o svijetu (tj. 'normalno' znanje, općeprihvaćeno konsenzusom). Za svaki pojam navedeni su ime (*Cyc® name*), definicija značenja, objašnjenje i uporaba pojma na engleskom te taksonomijske veze (*links*) kojima se pojmovi hijerarhijski uređuju i međusobno povezuju. Svaki je pojam u bazi znanja predstavljen pojmovnom natuknicom, ili osnovnom, konstantnom jedinicom (*constant, term, unit*) označenom simbolom #$. Natuknica može predstavljati *zbir* (*collection*, npr. skup ljudi), *pojedini predmet* (npr. neka osoba), *kvantitativ* (*quantifier, 'there exists'*) te *odnos* (*relation*: predikat, funkcija, prazno mjesto, atribut itd.) itd. Za svaku se natuknicu u tekstu/članku natuknice navode: *ime, definicija*, atributi *isa* (*'is an element of' one or more collections*), *genls* (*'subset of'*), hipertekstualne veze s nadređenim, općenitijim skupovima.

6.2. Mikrokosmos

http://crl.nmsu.edu/Research/Projects/mi ...intro-page.html

6.3. The Sensus ontology

http://ariadne.isi.edu:8005/sensus-edc/

6.4. The Organon — EASG Dictionary

http://www.gibson-design.com/philosophy/organon-dictionary-introduction.html

Program za ontološki konceptualni leksikon – filozofski pojmovnik.

7. EKSPERIMENTALNE LEKSIČKE BAZE ZA OBRADU PRIRODNIH JEZIKA

7.1. The Core Lexical Engine

http://www.cs.brandeis.edu/~rllc/corelex.html

Cilj ovog projekta jest kontekstualno određivanje značenja riječi. Za to je potrebno izraditi: generator osnovnog rječnika (*Core Lexical Engine*) i program za usvajanje novih riječi u određenu kontekstu, a uz pomoć statističkih metoda obrade korpusa.

Osnovni rječnik sadrži: osnovni semantički sadržaj, semantičke odnose za sve kategorije, generativne mehanizme za proširivanje i identifikaciju značenja riječi u kontekstu. Projekt razvija vlastiti model leksičke baze znanja dostupne automatskoj obradi, a utemeljen je na najnovijim modelima i teorijama generativnog leksikona (*Generative Lexicon Theory*, Pustejovsky, 1991; 1995).

7.2. Acquilex

www.ub.es/ling/acquing.htm

8. TEZAURUSI (pojmovnici)

8.1. Rogetovi tezaurusi

http://ecco.bsee.swin.edu.au/text/roget/

http://www.thesaurus.com/Roget-Alpha-Index.html

8.2. National Monuments Record

www.rchme.gov.uk/thesaurus/mar_place/c147823.htm

http://www.english-heritage.org.uk/index .html

www.english-heritage.org.uk/knowl...d199/index.html

Tezaurusi se koriste i kao strukturiran popis riječi sa svrhom standardizacije terminologije. *English Heritage* ima vodeću ulogu u standardizaciji terminologije. Rezultati standardizacije dani su na uvid korisnicima radi korištenja on-line i radi eventualnih primjedaba.

8.3. EUROVOC (Thesaurus Eurovoc Volume 2 Subject-oriented version)

http://www.europa.eu.int/celex/eurovoc

8.4. Hrvatski prijevod: EUROVOC, 2. svezak: Predmetna verzija:

izdala Hrvatska informacijsko-dokumentacijska referalna agencija – HIDRA, ZAGREB, 2000.

http://kuna.hidra.hr/eurovoc/index.htm

Pojmovnik sadrži 21 šire područje (politika, međunarodni odnosi Europske zajednice, pravo, ekonomija, trgovina, financije, društvena pitanja, obrazovanje i komunikacije, znanost, poslovanje i konkurencija, zapošljavanje i radni uvjeti, prijevoz, okoliš, poljoprivreda, šumarstvo i ribarstvo, poljoprivredno-prehrambena industrija, proizvodnja, tehnologija i istraživanje, energija, industrija, zemljopis, međunarodne organizacije) i 127 potpodručja koja obuhvaćaju sve djelatnosti Europske zajednice u obliku strukturiranog i kontroliranog popisa naziva za sve relevantne pojmove koji se javljaju u dokumentima Europskih zajednica. HIDRA je prevela 2. svezak (predmetna verzija) posljednjeg, trećeg izdanja iz 1995. godine, za potrebe sadržajne obrade službene dokumentacije Republike Hrvatske i povezivanja te dokumentacije s onom zemalja Europske unije radi jednoznačnijeg i preciznijeg prepoznavanja i komunikacije. Pojmovnik Eurovoc u izvornom je obliku objavljen na devet službenih europskih jezika Europske unije.

European Education Thesaurus[4] (dokumentacijski jezik namijenjen obradi višejezičnih podataka o obrazovanju u Europi, posebno za predmetno označavanje na području odgoja i obrazovanja.

8.5. Controlled Vocabularies

www2.fit.qut.edu.au/infoSys/middle/cont_voc.html

To je www stranica s najvećim popisom stručnih tezaurusa i drugim klasifikacijskim programima za rad s leksičkim bazama podataka.

REFERENCIJE

- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.
- Ellis, H. C. and Hunt, R. R. (1972/1993). *Fundamentals of Cognitive Psychology*. McGraw Hill.
- Hearst, M. (1998). *Categories and Attributes. www.sims.berkeley.edu.*
- Miller, G. A., Beckwith, R., and Fellbaum, C. (1993). *Five Papers. www.cogsci.princeton.edu~wn/Publications.*
- Pustejovsky, J. (1991). The Generative Lexicon. *Computational Linguistics*, 17(4).
- Pustejovsky, J.(1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Sperber, D. and Wilson, D. (1986) *Relevance: Communication and Cognition*. Oxford: Blackwell.

LEXICAL DATABASES ON THE INTERNET

Summary

A survey of lexical databases (LDB) is first presented including some theoretical issues of their definition and classification. A selected number of databases is shown and accompanied by the sources of lexical databases on the internet.

Key words: lexical databases, Internet

BILJEŠKE

[1] database, izg.: /'deitabeis/, tradicionalno u UK /'da:tabeis/

[2] *'Meaning is such a pervasive aspect of linguistic knowledge that a totally unbiased and perfectly balanced survey of theoretical and applied work involving lexical notions is a very difficult, perhaps impossible, project to carry out.' (EAGLES, 1999:4).*

[3] *morphosyntactic specifications (language independent and dependent); specifications for syntactic and semantic sub-categorisation, and common specifications for lexical data in corpora and lexicons.(EAGLES)*

[4] Hrvatska verzija ovog pojmovnika objavljena je u nas pod nazivom Europski prosvjetni pojmovnik, ur. M. Bratanić, priredili M. Bratanić *et al.*, Ministarstvo kulture, Zagreb, 1996.