

Navodila za popraviljanje slovenskega wordneta

različica 1.2

23.04.2009

Projekt

Jezikoslovno označevanje slovenskega jezika

<http://nl.ijs.si/jos/>

Odsek za tehnologije znanja

Institut Jožef Stefan

Pripravila: Darja Fišer

Kazalo

1	Cilj projekta	3
2	Osnovni pojmi.....	3
3	DEBVisDic.....	4
3.1	Navodila za namestitev urejevalnika DEBVisDic.....	4
3.2	Navodila za uporabo urejevalnika DEBVisDic	7
3.3	Primerjava sinsetov v obeh jezikih.....	7
3.4	Različni pogledi	8
3.4.1	Pogled Edit.....	9
4	Navodila za popravljanje wordneta	9
4.1	Viri	9
4.2	Večpomenski in večbesedni literali.....	9
4.3	Popravljanje sinseta	10
4.4	Dodajanje nadpomenk	11
4.5	Dodajanje literalov	11
5	Seznam koristnih spletnih virov.....	12

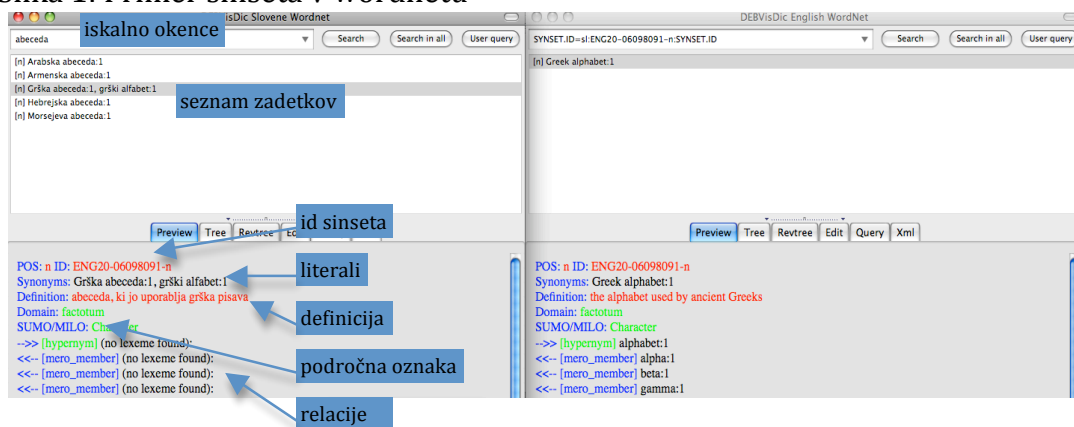
1 Cilj projekta

- pregled in popravljajanje wordneta glede na korpusne podatke

2 Osnovni pojmi

Wordnet je pojmovno zasnovan semantični leksikon, v katerem so pojmi glede na pomen med seboj povezani s semantičnimi relacijami. Pojmi v wordnetu so razdeljeni na samostalniške, glagolske, pridevniške in prislovne. Vse besede, ki označujejo isti pojem in imajo isti pomen, so povezane v **sinset**. Vsak sinset ima svoj **ID**. Sopomenke v sinsetu imenujemo **literali**, ki so lahko eno- ali večbesedni. Poleg sopomenske relacije so v wordnetu označene še nad- in podpomenke, protipomenke ter holo- in meronimi. Sinset vsebuje tudi definicijo, primere rabe, področno oznako (npr. avtomobilizem) in povezavo na ontologijo. Primer sinseta prikazuje Slika 1.

Slika 1: Primer sinseta v wordnetu



Dodatne informacije o slovenskem wordnetu najdete na projektni spletni strani:

<http://lojze.lugos.si/~darja/slownet.html>.

3 DEBVisDic

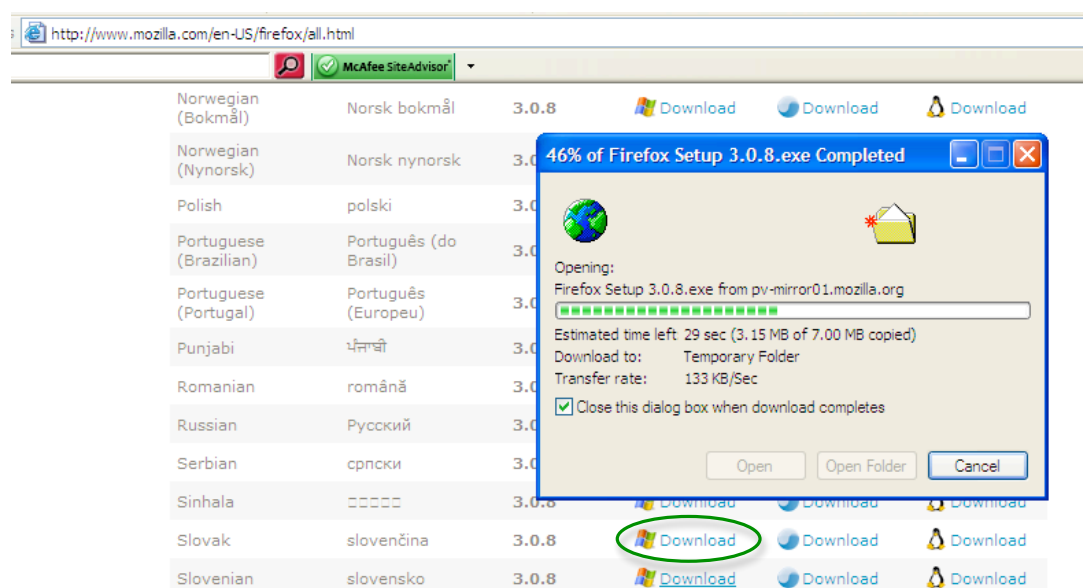
DEBVisDic je orodje za pregledovanje wordneta in ga potrebujete za sodelovanje pri projektu. V njem boste lahko iskali in popravljali sinsete v avtomatsko izdelanem wordnetu za slovenščino in ga hkrati primerjali z angleškim wordnetom. DEBVisDic si morate najprej namestiti na svoj računalnik, nato pa preko interneta dostopate do slovenskega in angleškega wordneta.

3.1 Navodila za namestitev urejevalnika DEBVisDic

DEBVisDic deluje v operacijskem sistemu Windows. Za namestitev DEBVisDica boste potrebovali spletni brskalnik Firefox 3, ki ga dobite tukaj:

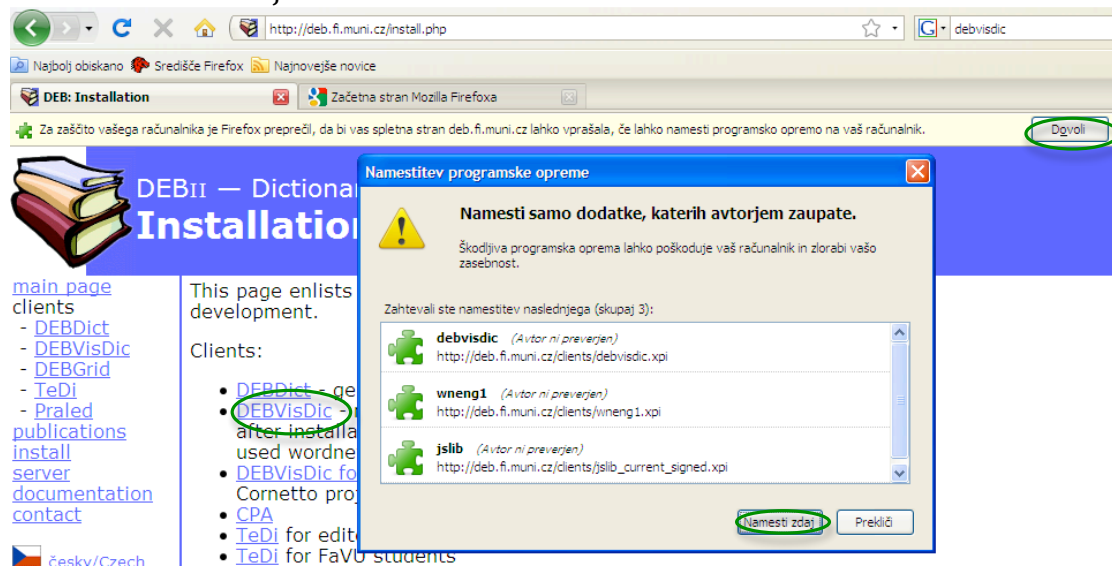
<http://www.mozilla.com/en-US/firefox/firefox.html> (glej Sliko 1).

Slika 1: Namestitev urejevalnika Firefox 3.0.8



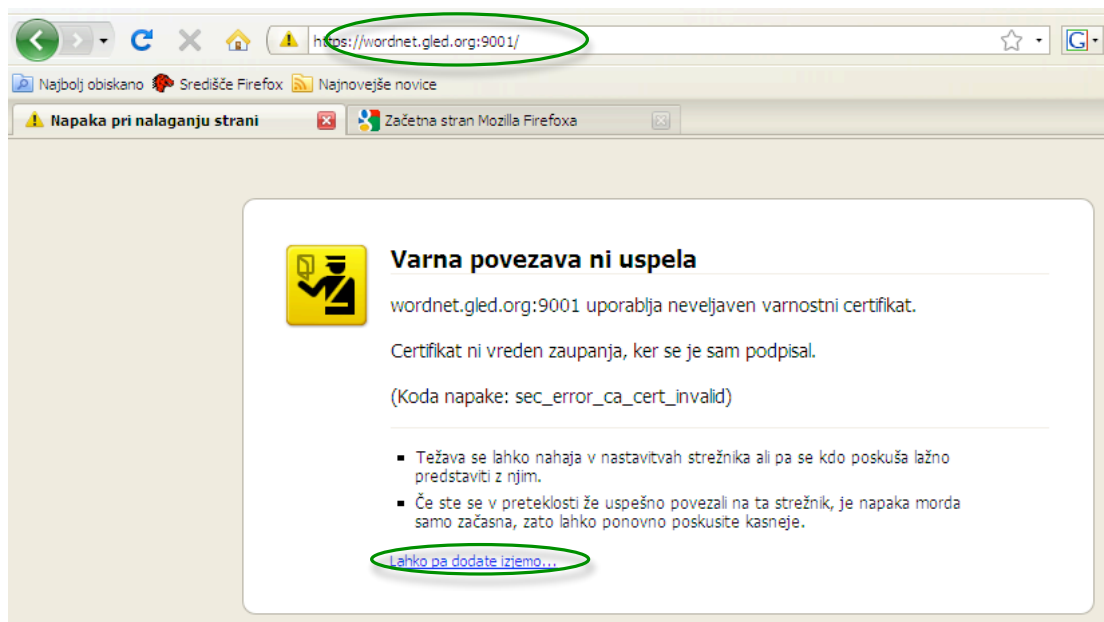
Ko ste uspešno namestili Firefox, si morate namestiti še DEBVisDic klient, ki ga dobite tukaj: <http://deb.fi.muni.cz/install.php> (druga povezava na strani, glej Sliko 2).

Slika 2: Namestitev urejevalnika DEBVisDic

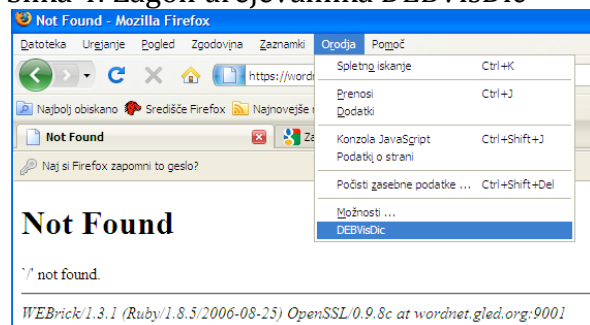


Zdaj ponovno zaženite Firefox. Pojdite na naslov <https://wordnet.gled.org:9001> in potrdite izjemo za certifikat (glej Sliko 3). Nato zaženite DEBVisDic (Tools → DEBVisDic, glej Sliko 4).

Slika 3: Dodajanje izjeme za certifikat

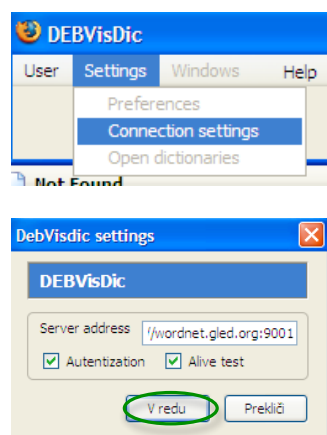


Slika 4: Zagon urejevalnika DEBVisDic



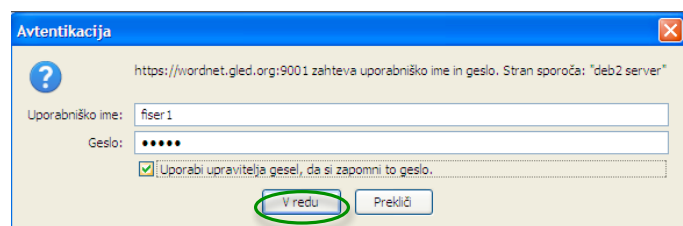
Ko ste DEBVisDic zagnali, se vam odpre prijavno okno, vendar se sprva ne boste mogli prijaviti, zato zaprite prijavno okno in spremenite naslov strežnika za DEBVisDic. Pojdite na Settings → Connection Settings in v okence Server Address vpišite <https://wordnet.gled.org:9001>. Ostalih nastavitev ne spreminjajte in kliknite OK (glej Sliko 5).

Slika 5: Nastavitev strežnika za dostop do wordneta



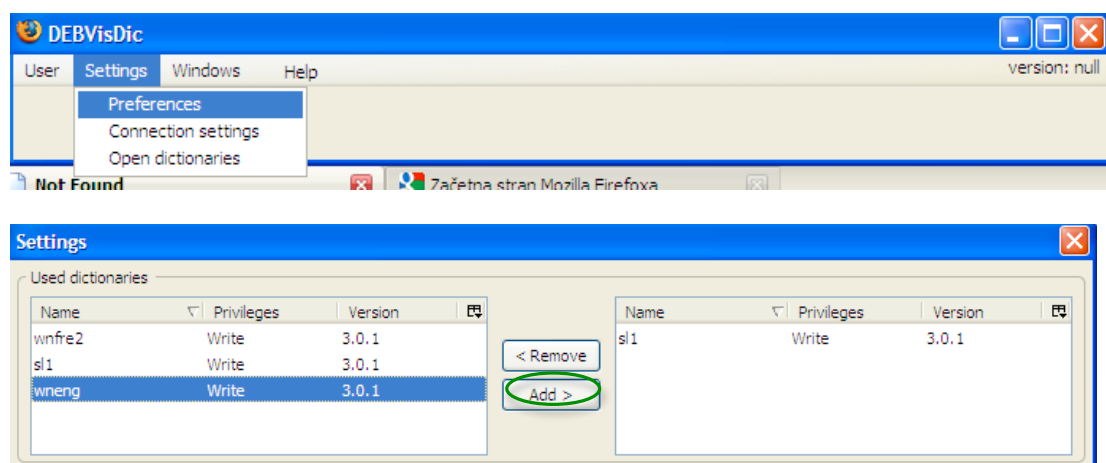
Zdaj v prijavno okno vpišite svoje uporabniško ime in geslo, kot prikazuje Slika 6.

Slika 6: Prijava na strežnik



Ko ste se uspešno prijavili na strežnik, pojdite na Settings → Preferences in izberite slovenski in angleški wordnet ter ju dodajte s klikom na gumb Add. Nato kliknite OK (glej Sliko 7).

Slika 7: Dodajanje slovenskega in angleškega wordneta



3.2 Navodila za uporabo urejevalnika DEBVisDic

Podrobnejša navodila za delo z DEBVisDicom najdete tukaj:

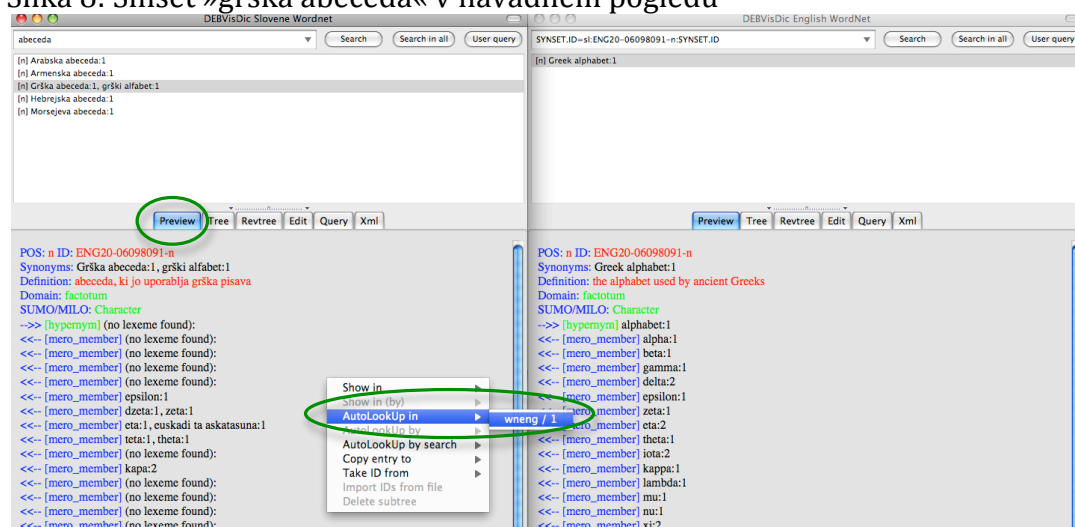
<http://nlp.fi.muni.cz/trac/deb2/wiki/DebVisDicManual> (v angleščini) in jih toplo priporočam. Na hitro pa tole:

3.3 Primerjava sinsetov v obeh jezikih

DEBVisDic v enem okencu prikazuje angleški, v drugem pa slovenski wordnet.

Iskano besedo vpišite v iskalno polje in kliknite gumb Search. Če želite, da se vam isti sinset prikaže tudi v angleškem wordnetu, z desno miško kliknite na prikazan sinset v največjem spodnjem oknu, in izberite ukaz »AutoLookUp in« (glej Sliko 8). Angleški wordnet uporabljajte za preverjanje ustreznosti slovenskih prevodov (upoštevajte sinonime, definicijo, primere rabe in semantične relacije).

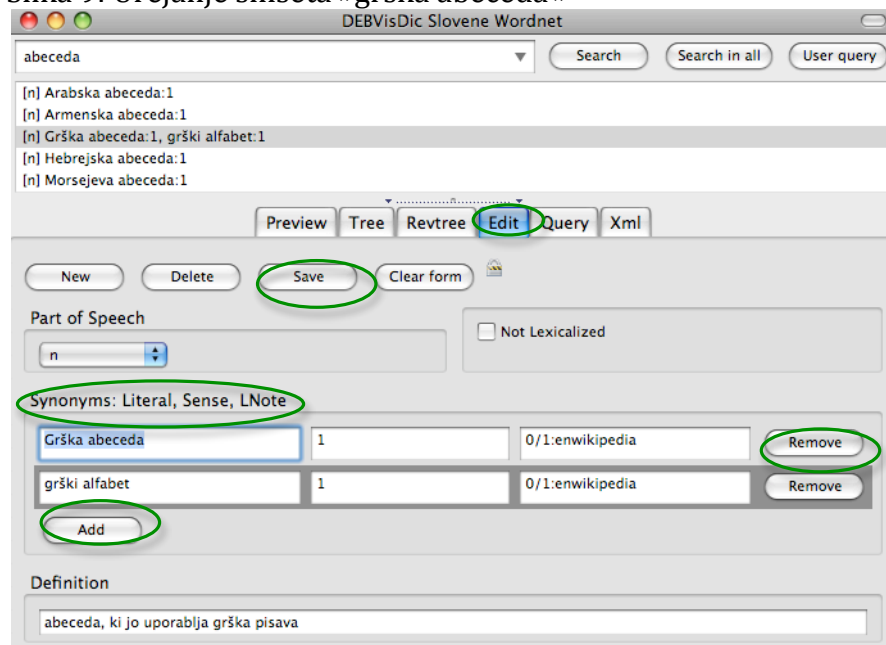
Slika 8: Sinset »grška abeceda« v navadnem pogledu



3.4 Različni pogledi

V zavihku »Preview« se prikaže celoten sinset. V zavihku »Tree« si lahko ogledate celotno drevo, v katerega sinset sodi. V zavihku »Edit« pa popravljate napake in dodajate manjkajoče sinonime (glej Sliko 9).

Slika 9: Urejanje sinseta »grška abeceda«



3.4.1 Pogled Edit

Vse sinonime prikažete s klikom na plus v kategoriji »Synonyms: Literal, Sense, LNote«. V levem okvirčku so posamezni sinonimi, v naslednjem okvirčku št. pomena, v zadnjem pa vir, iz katerega je bil literal pridobljen. Na koncu vrstice je gumb »Remove«, na dnu kategorije pa gumb »Add«.

Gumb »Remove« uporabite, kadar je celoten sinonim neustrezen in ga želite odstraniti. Gumb »Add« uporabite, kadar pomemben sinonim v sinsetu manjka in ga želite dodati. Ko želite spremeniti le obstoječ sinonim, spremembe vnašate v levi okvirček (»Literal«). Spremembe shranite s pritiskom na gumb »Save« na vrhu okna (glej Primer 1).

Primer 1: V sinsetu »grška abeceda« je prvi literal napačno zapisan z veliko začetnico, ki jo moramo popraviti v malo. Drugi literal se nam zdi nenavaden, zato ga preverimo v korpusu FidaPLUS, ki vrne 61 zadetkov, pa še v SSKJ ga najdemo, zato je literal v wordnetu ustrezen in ga ne spreminjamo. Definicija sinseta je v slovenščini, vendar je pravilna, zato tudi te ne spreminjamo. Nato kliknemo še gumb »Save«.

4 Navodila za popraviljanje wordneta

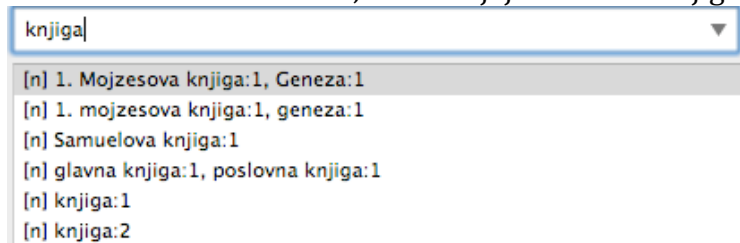
4.1 Viri

Pri projektu uporabljajte vse eno- in večjezične slovarske, glosarske in korpusne vire, ki so za slovenščino na voljo: SSKJ, FidaPLUS, Evroterm itd (glej poglavje 6). Svoje predloge za prevode vedno preverite v virih in si jih in njihovih pomenov ne zmišljajte.

4.2 Večpomenski in večbesedni literali

Isti literal v wordnetu se lahko pojavi v enem samem sinsetu (enopomenska beseda oz. besedna zveza) ali v več kot enem sinsetu (večpomenska beseda oz. besedna zveza). Vaša naloga je, da pregledate **vse pomene besede**, ki vam je bila dodeljena, pa tudi **vse pomene večbesednih zvez**, ki vašo besedo vsebujejo (glej Sliko 10 in Primer 2).

Slika 10: Seznam sinsetov, ki vsebujejo besedo »knjiga«



Primer 2: Zanima nas beseda »knjiga«, ki jo vpišemo v iskalno okence. Med zadetki se pojavita dva pomena besede »knjiga«, poleg tega pa tudi štirje večbesedni izrazi, ki vsebujejo besedno »knjiga«, zato preverimo vseh šest sinsetov in v njih odpravimo morebitne napake (glej Primer 3).

4.3 Popravljanje sinseta

Slovenske sinsete primerjajte z angleškimi. Če v slovenskem wordnetu opazite kakšno napako, pojdite na jeziček Edit in jo popravite oz. izbrišite napačen literal v sinsetu. Pri tem upoštevajte angleške sinonime za ta pomen, definicijo in primere rabe ter semantične relacije sinseta (glej Primer 3).

Primer 3: Pri pregledovanju sinsetov, ki vsebujejo besedo »knjiga«, opazimo, da prva dva sinseta vsebujeta enake literale. Ko sinseta preverimo še v angleškem wordnetu, ugotovimo, da sta literala ustrezna za prvi pomen (the first book of the Old Testament), za drugega pa je ustrezen samo literal »geneza« (genesis:1, generation:5, a coming into being), zato napačen literal iz tega sinseta izbrišemo. Prav tako moramo popraviti sinset »Samuelova knjiga« v »Samuel«. Ostali sinseti so pravilni in jih ne spreminjamo.

V angleškem wordnetu ne popravljajte ničesar. V slovenskem wordnetu pa **popravljajte samo literale** (sinonime), primerov rabe in semantičnih relacij ne spreminjajte, definicije pa samo, če so v slovenščini in v njih opazite kakšno napako. V tem primeru napačne slovenske definicije zamenjajte z angleško definicijo, ki jo najdete pri istem sinsetu v angleškem wordnetu (glej Primer 4).

Primer 4: Pri pregledovanju sinseta »Križev pot« opazimo, da vsebuje naslednjo definicijo: poleg tega tudi molitev, ki jo verniki molijo, ko se spominjajo tega trpljenja. Ker ji manjka prvi del definicije, jo ali izpolnimo, če dela ni veliko, ali zamenjamo z angleško definicijo iz angleškega sinseta.

4.4 Dodajanje nadpomenk

Pri pregledovanju wordneta pogledite tudi **vse nadpomenke** dodeljene besede do vrha taksonomije. Če nadpomenka še ni prevedena v slovenščino (»no lexeme found«), s klikom na gumb Go To pojdite nanjo in vanjo dodajte ustrezen prevod oz. prevode (glej Sliko 11 in Primer 5).

Slika 11: Nadpomenka sinseta »grška abeceda« je prazna

POS: n ID: ENG20-06098091-n
Synonyms: Grška abeceda:1, grški alfabet:1
Definition: abeceda, ki jo uporablja grška pisava
Domain: factotum
SUMO/MILO: Character
-->> [hypernym] (no lexeme found):

Relations

Relation	Value	Action
[*] [n] (no lexeme found):	hypernym	Go to Remove

Add

Primer 5: Pri pregledovanju sinseta »grška abeceda« opazimo, da je njegova nadpomenka prazna (»no lexeme found«) Zato se v pogledu za urejanje premaknemo na ta sinset s klikom na gumb »Go to«, ga preverimo v angleškem sinsetu (alphabet:1, a character set that includes letters and is used to write a language) in ustrezen prevod »abeceda« vstavimo kot literal za ta sinset v slovenski wordnet.

4.5 Dodajanje literalov

Če v wordnet dodate kak literal, ga vpišite v okence Literal. Okence Sense v angleškem delu wordneta vsebuje številko pomena večbesednih literalov, v slovenščini pa ga uporabljamo za štetje pojavitev v korpusu pri avtomatski izdelavi wordneta. Vi v to okence vpišite črko »x«, v okence LNote pa vir, v katerem ste slovenski literal našli. Če želite dodati več sopomenk, vsako sopomenko vpišite v novo vrstico kot nov literal, ne v isti okvirček, ločen z vejicami. Spremembe shranite s klikom na gumb Save (glej Sliko 12).

Slika 12: Primer dodajanja literala »abeceda«

Synonyms: Literal, Sense, LNote

Literal	Sense	LNote	Action
abeceda	x	sskj	Remove

Add

5 Seznam koristnih spletnih virov

Wordnet:

- Princeton WordNet: <http://wordnet.princeton.edu/>
- SloWNet: <http://lojze.lugos.si/~darja/slownet.html>

Programska oprema:

- Firefox 3.0: <http://www.mozilla.com/en-US/firefox/firefox.html>
- DEBVisDic: <http://deb.fi.muni.cz/install.php>

Slovarji in glosarji:

- SSKJ: <http://bos.zrc-sazu.si/sskj.html>
- Pravopis
- Evroterm: <http://evroterm.gov.si/index.php?jezik=angl>
- Islovar : http://www.islovar.org/iskanje_enostavno.asp

Korpusi:

- FidaPlus: <http://www.fidaplus.net/>
- Sketch Engine: <http://www.sketchengine.co.uk/>
- korpusi na IJS: <http://nl2.ijs.si/>