Miran Željko
Ljubljana

UDK 81'374:004.9:81'25

IZBOLJŠAVE ELEKTRONSKIH SLOVARJEV[1]

Članek obravnava možnosti za izboljšavo slovarjev z vidika prevajalca. Prevajalcem so slovarji osnovni pripomočki pri delu, v zadnjem času se jim pridružujejo tudi dvojezični korpusi. Obstoječi slovenski elektronski slovarji so nastali iz izdelkov, ki so bili prvotno narejeni za knjižno obliko, nato pa so bili ti slovarji po liniji najmanjšega odpora predelani v računalniške programe. Elektronski slovarji bi lahko omogočali bistveno več kot knjižni slovarji: iskanje po celotnem besedišču v slovarju, iskanje podobnih besed, uporabo korpusa za primere rabe, dinamično povezavo med slovarjem in korpusom, stalno dopolnjevanje slovarja namesto priprave popolnoma novih slovarjev vsakih nekaj desetletij. Kot primer praktične uporabe predlaganih načel je na kratko predstavljena terminološka zbirka *Evroterm*.

terminološka zbirka, spletni slovar, korpus, stalne izboljšave, iskanje po polnem besedilu, *Evroterm*

The article presents some possibilities for improving dictionaries from the translator's point of view. Dictionaries, glossaries and terminology databases (as well as, recently, parallel corpora) are the basic tools for translators. The existing Slovene electronic dictionaries are based on dictionaries in book form – the data from those books were transformed into computer software following the line of least resistance. However, electronic dictionaries could provide more functionality than books: full-text search, fuzzy search, corpora as a source of collocations, dynamically linked dictionary and corpus, and continuous improvements of a dictionary instead of new dictionary projects every few decades. As an example of practical use of the proposed improvements, the *Evroterm* terminology database is presented.

terminology database, on-line dictionary, corpus, continuous improvements, full-text search, *Evroterm*

1 Uvod

Če razvoj slovarjev primerjamo z razvojem avtomobilov, smo trenutno v tisti fazi, ko je Carl Benz izpregel konje izpred kočije in nanjo namestil motor, zdaj pa se ponosno prevaža naokoli. Slovarji danes v Sloveniji praviloma najprej izidejo v knjigi, potem pa jih računalnikarji pretvorijo v elektronsko obliko (Anžlovar 2004) – najprej torej izdelamo kočijo, potem odžagamo oje in kočijo predelamo tako, da je nanjo možno namestiti motor.

[1] V tem prispevku uporabljam izraz »slovar« za vse oblike slovarja in slovarjem podobnih pripomočkov – torej tudi za glosarje in terminološke zbirke.

V prihodnosti bo treba postopek povsem spremeniti: kočija in avto sta dva različna izdelka, zato ju je treba snovati in izdelovati ločeno; slovarji v knjižni obliki naj bodo za bibliofile, prevajalci pa večinoma potrebujejo slovarje v elektronski obliki in pri sestavljanju takih slovarjev je treba čim bolj izkoristiti vse možnosti informatike. Potem se ne bo dogajalo to, da po angleško-slovenskem slovarju ne moremo iskati slovenskih besed (npr. *Veliki angleško-slovenski slovar* na CD-ju) in bo preprosto graditi večjezične slovarje. Primeri rabe, kolokacije in izjeme ne bodo omejeni na znanje ali prepričanje sestavljavcev slovarja, temveč jih bo slovar samodejno potegnil iz korpusa.

2 Izboljšave slovarjev

2.1 Iskanje po celotnem besedilu

Prvi elektronski slovarji so bili knjige, pretvorjene v elektronsko obliko – taki so pri nas npr. Amebisovi slovarji (http://www.amebis.si): vsebina elektronske verzije *Velikega angleško-slovenskega slovarja* je enaka vsebini knjige s tem naslovom, le iskanje je hitrejše, ker ni treba obračati strani. Večja preglednost je dosežena z uporabo različnih barv za iztočnico, prevode in besedne zveze, možno je iskati v polju zadetkov, v poljubnem strokovnem področju in po primerih rabe, možno je dodajati opombe – in ob še nekaterih manjših dopolnitvah se tu napredek praktično konča.

Manjka pa najpomembnejša izboljšava – čeprav bi jo bilo možno narediti povsem preprosto: **namesto da je iskanje omejeno samo na angleške iztočnice, bi lahko iskali tudi po slovenskih prevodih besed in besednih zvez**. Programerji znajo rešiti to nalogo, a najbrž založnik meni, da bi imel slovar s tem večjo vrednost, kot so mu jo bili pripravljeni dodeliti, saj bi naenkrat postal tudi slovensko-angleški slovar. Dva slovarja za ceno enega na trgu, kjer skoraj ni konkurence – to pa res nima smisla!

2.2 Iskanje podobnih besed

Pri pisanju občasno prihaja do napak; lahko zaradi tipkanja, ali pa, ker si človek napačno zapomni zapis besede. Med pisanjem v urejevalniku besedil nas črkovalnik opozori na morebitno napako; če želimo, nam predlaga (po njegovem mnenju) pravilno besedo. Pri iskanju besede v slovarju pa v takem primeru sploh ne dobimo zadetka. **Če elektronski slovar ne najde besede, ki jo vtipkamo, bi bilo prijazno do uporabnika, če bi program prikazal tudi zadetke, ki so podobni iskani besedi.**

2.3 Korpusi

Pripomoček, katerega korist prevajalci spoznavajo šele v zadnjih letih, so korpusi prevodov. Uvod v korpuse je npr. v Hirci 1999.

Slovar in korpus se zdita povsem različna izdelka: v slovarju so podatki urejeni po abecedi, korpus pa je neurejena zbirka in šele ob izpisu rezultatov iskanja dobimo iz nje urejen izvleček. V resnici je podobnost med slovarjem in korpusom precej večja, kot se zdi na prvi pogled: če kot najosnovnejšo obliko slovarja vzamemo **glosar, v katerem vsaki besedi v prvem jeziku ustreza beseda v drugem jeziku, je** to **najenostavnejša oblika korpusa**. Po drugi strani pa bi v dovolj velikem dvojezičnem korpusu našli vse besede iz slovarja, le preslikave med besedami moramo poiskati sami – več o tem je npr. v Vintar 2003 – **korpus torej lahko obravnavamo kot neke vrste neurejen glosar ali glosar z veliko količino šuma**.

Ponavadi se korpusi in slovarji obravnavajo ločeno – na spletu je na voljo npr. *Slovar slovenskega knjižnega jezika* in korpus *Nova beseda*, ki vsebuje tudi slovensko leposlovje. Smiselno bi bilo, da bi bile ob zadetkih iz *Slovarja slovenskega knjižnega jezika* narejene povezave na primere iz korpusa slovenskega leposlovja – tako bi videli, kako se iskana beseda uporablja v knjižnem jeziku – namesto tega pa je na spletu ista verzija slovarja z enakimi primeri, kot so jih v knjižno obliko vnesli sestavljavci slovarja. Enako je s CD-jem.

Razumljivo je, da je v knjižni obliki slovarja število primerov rabe omejeno, saj smo omejeni z naravo medija – odvisno od debeline papirja je možno knjigo kolikor toliko neproblematično uporabljati, če obsega do približno 2000 strani. Če je slovar preobširen, postane pretežek. Ta problem rešimo tako, da slovar izide v več zvezkih – a tudi tu se pri praktični rabi hitro pojavi meja. Pri računalniških medijih je ta omejitev nekaj redov velikosti višja – ste poskusili izračunati, koliko znakov je v *Velikem angleško-slovenskem slovarju* (Grad, Škerlj, Vitorovič 1997). V knjigi je skoraj 1400 strani, na strani je 5000 do 6000 znakov in če ta podatka zmnožimo, dobimo med 7 in 8,4 milijonov znakov, kar je le 1 % kapacitete CD-ja!

Ena od osnovnih umetnosti sestavljanja slovarjev je torej tudi izbira ustreznih primerov rabe (več o tem je v Drstvenšek 2003). Pri tem lahko nastane več težav:

– vsak avtor ima omejeno znanje in nekaterih primerov ne vključi v slovar (pogosto izpadejo novejše besedne zveze – prav te pa bi bile za uporabnike slovarja najzanimivejše);

– morda poskuša avtor dokazati kako svojo hipotezo in izbere tiste primere, ki potrjujejo njegovo mnenje, nasprotnih pa ne uvrsti v slovar;

– avtorji slovarjev so običajno ljudje z večdesetletnimi izkušnjami – zato se v slovarju znajdejo tudi besede in besedne zveze, ki jih v sodobni rabi redkeje srečamo.

Te probleme rešimo, če sestavimo korpus, in slovar priredimo tako, da išče primere rabe neposredno v korpusu – če je korpus sestavljen uravnoteženo in če kakšni primeri niso namenoma odstranjeni, bi morali dobiti dejanske primere rabe. Dejstvo je, da zaradi velike količine podatkov v korpusu prihaja do napak, vendar iz množice zadetkov – kljub morebitnim napakam – običajno lahko izluščimo pravilo.

2.4 Korpus v povezavi s slovarjem

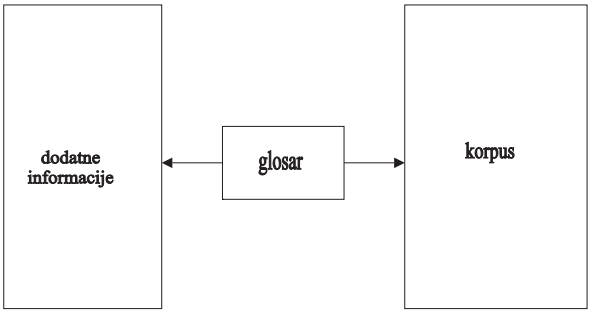
Iz knjižne izdaje slovarjev smo navajeni, da so primeri rabe navedeni statično – v knjigah ne more biti drugačnega načina prikaza. Pri elektronskem slovarju pa je smiselno, da je povezava med iztočnico in primeri rabe dinamična – vzpostavi se šele pri iskanju.

Z vidika prevajalca lahko na splošno rečemo, da posamezen zapis v slovarju sestavljajo trije deli (slika 1):

– **glosar** (prevod iztočnice v prvem jeziku v besedo v drugem jeziku),

– **dodatne informacije** o iztočnici (odvisno od besedne vrste, jezika, obsega slovarja in ciljnih uporabnikov); seznam podatkov, ki naj bi jih vsebovala terminološka zbirka, je naštet v standardu *ISO 12616*;

– **primeri rabe** – tu je najenostavneje uporabiti kar korpus prevodov; tudi tu lahko navedemo dodatne podatke: zanesljivost prevoda, področje, vir, celotno besedilo ipd.



Slika 1: Posamezni elementi slovarja in povezave med njimi

V takem sistemu imamo lahko dve smeri iskanja:

1. *glosar* → *dodatne informacije* → *korpus*: uporabnik običajno ne potrebuje vseh podatkov, zato je smiselno, da se mu informacije odpirajo postopno: najprej iz glosarja dobi seznam iztočnic (lahko tudi s prevodi), ki ustrezajo iskanemu kriteriju, ob kliku izbrane besede dobi dodatne informacije, ob ponovnem kliku pa primere rabe (seznam zadetkov iz korpusa z besedili v izvornem in ciljnem jeziku). Obstajajo še druge možnosti, način izpisa pa je odvisen od izvedbe slovarja (v slovarju, ki je nameščen na računalniku uporabnika, se lahko izvajajo zahtevnejše operacije kot v slovarju, ki je na spletnem strežniku), namena slovarja, količine informacij in od predvidene običajne poti iskanja.

2. *glosar* → *korpus* → *dodatne informacije*: v slovarjih je vedno omejena množica besed; če je slovar splošen, v njem manjkajo strokovni izrazi, če je slovar omejen na neko strokovno področje, pa v njem ni splošnih izrazov, zato se vedno zgodi, da kakega pojma ne najdemo v slovarju. Ob primerno velikem korpusu pa je zelo verjetno, da je vsaj nekaj iskanih pojmov v korpusu, zato je smiselno iskanje omogočiti tudi v drugi smeri, pri čemer se uporabniku informacije spet odpirajo

postopno: najprej se iskana beseda poišče v glosarju, nato pa še v korpusu in se iz okolice iskane besede najde pomen neznane besede. Če iskane besede ni v glosarju, izpišemo seznam zadetkov iz korpusa. Če pa je iskana beseda tudi v glosarju, je smiselno, da ob njej uporabniku ponudimo tudi pot do dodatnih informacij o pojmu.

Z vidika prevajalca ima povezava slovarja in korpusa naslednje **prednosti**:

– več načinov iskanja,

– večjo verjetnost, da najde pomen iskane besede,

– več podatkov o iskani besedi,

– hitrejše in lažje iskanje.

Pomanjkljivost tega pristopa je predvsem v rabi korpusa. V dobrem korpusu je več deset milijonov besed. Vseh podatkov v korpusu ni možno preveriti (oziroma bi bilo to preverjanje predrago), zato so v korpusih napake; praviloma je v korpusih več napak kot v slovarjih. Uporabnike je treba opozoriti na to – če kak zadetek odstopa od drugih, je možno, da ne gre za izjemo, temveč za napako in v takem primeru mora uporabnik preveriti podatek še v kakem drugem viru. Pripombe uporabnikov o napakah so zelo dobrodošle in na osnovi teh povratnih informacij je treba korpus redno osveževati.

2.5 Prikaz na ekranu

Pri iskanju v slovarjih in korpusih običajno dobimo kot rezultat veliko količino podatkov. Te je treba urediti in na ekranu prikazati tako, da uporabnik čim prej najde iskano. **Osnovna pomoč pri tem so barve** – različne informacije naj bodo različno obarvane, manj pomembni podatki so preprosto v črni barvi. Pomoč barv pri uporabi slovarja je razvidna že iz Amebisovih slovarjev. Pri korpusu je uporabniku v pomoč, če so posamezne enote jasno ločene druga od druge in če je iskana beseda pobarvana (če je natisnjena samo krepko – kot je npr. v korpusu *SVEZ-IJS* – jo je na gosto popisanem ekranu težje najti). **Pri vzporednem dvojezičnem korpusu je lažje najti ustreznico v obeh jezikih, če sta izpisa v obeh jezikih vzporedna** (pri zaporednem izpisu zaradi daljše poti, ki jo mora opraviti oko, vzporejanje traja dalj časa), kar lahko vidimo, če primerjamo korpus *SVEZ-IJS* z *Evrokorpusom*. Dodatna pomoč uporabniku je, če je v izpisu zadetkov iz korpusa pobarvan tudi prevod iskane besede (če ga program najde v glosarju).

2.6 Stalne izboljšave

V preteklosti je bil način dela tak, da se je zbrala ekipa, naredila slovar in ta je (odvisno od kakovosti in pomembnosti) nespremenjen doživel nekaj ponatisov in bil naprodaj več let ali celo desetletij.

Pri elektronskih slovarjih je pristop lahko drugačen: osnovno verzijo slovarja naredimo podobno kot prej. V vsakem slovarju so napake in pomanjkljivosti, tudi če je narejen še tako skrbno. S spremembo knjižne verzije slovarja so veliki stroški (poleg očitnih stroškov s pripravo in tiskom nove izdaje se pojavi še nekaj vprašanj:

kdo bi kupil preostalo naklado stare izdaje z napakami, če je na voljo novejša različica; kako pogosto (pri kolikšnem številu napak) izdajati osvežene verzije ipd.). Pri elektronskih slovarjih so te težave rešene s samo naravo medija, saj je strošek za CD bistveno nižji od stroškov tiska knjige, še vedno pa so problemi, ker obstaja več različic slovarja. Vse se bistveno poenostavi z uporabo interneta: **če je slovar na spletu, moramo osveževati podatke le na enem mestu, vsak uporabnik pa ima v vsakem trenutku na voljo najnovejšo različico.** Ker ljudje nimajo vedno pri roki računalnika v povezavi z internetom, je smiselno omogočiti dostop do slovarja tudi z mobilnim telefonom.

V proizvodnji že več desetletij uporabljajo načelo *Demingovega kroga stalnih izboljšav* in to idejo lahko smiselno uporabimo tudi pri razvoju slovarjev.

2.6.1 Vsebinske izboljšave

Pri osveževanju gre za več opravil:

– popravljati je treba napake,
– dodajati je treba nove iztočnice,
– dodajati je treba nove pomene obstoječim iztočnicam,
– pojme, ki v tem procesu zastarijo, je treba ustrezno označiti.

Na najočitnejše napake nas opozorijo uporabniki slovarja, postavi pa se vprašanje, katere iztočnice dodajati v slovar. Glede tega imamo lahko različne pristope:

1. Če pri takem projektu sodeluje računalnikar, bo trdil, da se v računalništvu pojavlja največ novih besed, zato je najpomembneje dodajati te besede. Strokovnjaki iz drugih ved imajo lahko drugačne utemeljitve: pravnik bo trdil, da se slovenski in angleški pravni red že v osnovi bistveno razlikujeta in je zato treba v slovarju to razliko čim bolj osvetliti; predstavnik kake humanistične stroke bo morda trdil, da slovarja z njegovega področja sploh ni in mora biti zato več tovrstnega besedišča v splošnem slovarju. Razmerja med novimi besedami z različnih področij je pri takem pristopu zelo težko določiti.

2. Druga možnost je, da uporabimo čim večji korpus, izračunamo frekvenco pojavljanja besed in dodajamo predvsem besede, ki se večkrat pojavljajo. Pri tem postopku nekako avtomatiziramo prej omenjeni postopek, a zadeva deluje le, če se korpus stalno osvežuje – v starem korpusu ne moremo najti novih besed. Lönneker (2004) predlaga, da pri takem dopolnjevanju kot vir za nove besede uporabimo korpus literarnih virov.

3. Tretja možnost je najpreprostejša in najbolj demokratična, seznam besed, ki ga je treba dodati, pa se ustvarja kar sam: **dodajamo tiste besede, ki jih uporabniki niso našli v obstoječi verziji slovarja**. Vsak slovar je praviloma narejen zato, da bi ga uporabljali drugi ljudje, ne avtorji slovarja. V Jakopin 2004 je navedena možnost, da analiziramo dnevnik spletnega strežnika. Pogosto (vsakodnevno) osveževanje slovarja razen pri redkih izjemah (npr. dopolnjevanje terminološke

baze med prevajanjem zelo obsežnega besedila) ni izvedljivo zaradi prevelikih stroškov. Če pa se osveževanja lotimo le enkrat letno, lahko pri zelo obiskanih strežnikih nastanejo težave, ker se dnevniške datoteke izredno napihnejo (npr. na strežniku www.gov.si se vsakih 10 dni ustvari približno 1 GB dnevnika). Boljša rešitev je, da program, ki išče po slovarju, v datoteko sam zapisuje besede, ki jih ni v trenutno delujoči različici slovarja. Najpogostejše besede s tega seznama so prvi kandidati za dopolnitev slovarja.

Na to, da je treba pregledati podatke tudi z vidika zastarevanja besed, redkeje pomislimo – a je pri takem načinu dela treba biti pozoren tudi na to. S to temo se ukvarja npr. Brookes (2004).

Z rednim osveževanjem podatkov lahko precej odpravimo pomanjkljivosti zaradi napak v slovarju in korpusu.

2.6.2 Tehnične izboljšave

Tehnične postopke osveževanja (pretvorba zapisa, prenos podatkov med strežniki, statistične obdelave) je večinoma možno avtomatizirati. Pogostost osveževanja je odvisna od tega, koliko novih ali spremenjenih podatkov se pojavi v časovni enoti – osveževanje je lahko mesečno, tedensko ali celo dnevno.

Poleg vsebine slovarja lahko spreminjamo tudi funkcionalnost programa – nove funkcije so na voljo vsem uporabnikom od trenutka, ko jih uvedemo.

Res je, da se s tem pojavijo dodatni stroški, vendar je vrednost slovarja bistveno večja, saj v vsakem trenutku vsebuje ažurne informacije. Velike zagonske stroške imamo samo ob prvi pripravi slovarja.

Redno osveževanje podatkov uporablja npr. Telekom pri telefonskem imeniku: kmalu potem, ko priključijo novega naročnika, so njegovi podatki na voljo na spletu.

2.7 Zaščita intelektualne lastnine

V pripravo slovarjev je vloženega bistveno več dela kot npr. v pisanje romana, zato je prodajna cena ustrezno višja, zaradi tega pa se pogosteje pojavi problem nedovoljenega kopiranja. Slovarja v knjižni obliki nima smisla kopirati, saj bi bila cena tega enaka ali celo višja od cene v knjigarni, obenem pa je tak slovar težje uporabljati (trenutno zanemarimo dejstvo, da knjig brez privoljenja avtorja ali založnika sploh ni dovoljeno kopirati). Amebisovi slovarji na CD-jih so sicer zaščiteni tako, da jih je možno namestiti le na en disk, vendar je to zaščito možno obiti (opis postopka je na internetu). Če se držimo pravil, ki jih predpisuje založnik, pa zabredemo v drugačne težave:

– denimo, da imam namizni in prenosni računalnik, uporabljam pa le enega naenkrat: pri takem načinu zaščite z eno licenco za elektronski slovar ne pridem skozi;

– denimo, da se mi pokvari disk, na katerem je nameščen slovar, podatkov pa ni možno obnoviti in je treba disk zamenjati;

– ali še huje: denimo, da mi ukradejo računalnik.

V zadnjih dveh primerih je na zaščitni disketi zapisano, da sem program namestil na disk – a diska nimam več. Z računom in ustreznimi potrdili je od proizvajalca verjetno možno dobiti novo zaščitno disketo, je pa ob tem nekaj dodatnih opravkov in nekaj dni bom brez slovarja.

Do težav prihaja zato, ker je dovoljenje za uporabo programa omejeno na računalnik, namesto da bi bilo omejeno na osebo.

Če želimo še znižati stroške in slovarja na CD-ju sploh ne izdamo, temveč imamo vse podatke le na spletu, se zdi, da je zaščita še slabša kot pri CD-jih (zaščita z imenom in geslom ni resna zaščita, ker si ljudje posojajo gesla).

Vendar obstaja tudi možnost profesionalne zaščite – banke jo uporabljajo za stranke, ki želijo imeti dostop do svojih računov prek interneta, državna uprava pa uporablja ta postopek za komunikacijo z državljani pri prenosu zaupnih podatkov (npr. oddaja napovedi dohodnine) **– to so spletna potrdila**. Na naslovu http://www.sigen-ca.si je predstavitev agencije SIGEN-CA, ki izdaja spletna potrdila za državljane in poslovne subjekte, ter opisi postopkov za pridobitev spletnih potrdil in njihovo uporabo.

Postopek uporabe bi bil v grobem takle: organizacija, ki bi želela na tak način omejiti dostop do svojih slovarjev (ali drugih podatkov), bi morala pridobiti spletno potrdilo za svoj strežnik, uporabniki (kupci) slovarja pa bi morali dobiti spletno potrdilo za svoj brskalnik (to potrdilo lahko dobite brezplačno na upravnih enotah), možno pa bi bilo uporabiti tudi bančno spletno potrdilo (če ga potencialni kupec že ima, nima pa npr. potrdila SIGEN-CA). Kupec bi plačal letno naročnino, ki bi bila bistveno manjša od zneska za nakup slovarja, in bi potem imel za neko obdobje dostop do slovarja. Ob prijavi uporabnika na slovarski strežnik bi ta zahteval spletno potrdilo uporabnika in bi ga preveril s stanjem v svoji bazi (do kdaj ima dotični uporabnik dovoljenje za dostop do strežnika). Po preverjanju podatkov bi uporabnik delal kot običajno. Podrobnosti glede dogajanja na strežniku so na omenjeni spletni strani (Poslovni subjekti – spletna potrdila (strežniki) → uporaba spletnega potrdila → izberemo vrsto strežnika).

Podobno kot pri posojanju gesel je tudi tu možno, da si uporabniki medsebojno izmenjajo spletna potrdila – a je ta možnost bolj teoretična, saj bi imel lastnik sposojenega bančnega spletnega potrdila s tem možnost opravljanja vseh bančnih storitev v imenu druge osebe (kar je podobno, kot če nekomu posodimo bančno kartico in mu zaupamo tudi geslo za bankomat) ali dostop do vseh zaupnih podatkov pri poslovanju z državno upravo (pri uporabi spletnega potrdila SIGEN-CA) – kar je celo več, kot če bi nekomu posodili osebno izkaznico, ker pri spletnem potrdilu ni možno preveriti videza lastnika. Občasno bi morda prišlo do zlorab dostopa, a menim, da bi bilo teh pojavov bistveno manj, kot je nedovoljenega kopiranja CD-jev.

Dobrih strani prenosa slovarja na splet in zaščite s spletnimi potrdili je več:

– ponudnik slovarja vzdržuje podatke na enem mestu;

– vsi uporabniki imajo dostop do zadnje verzije slovarja;

– ni več izdelave in distribucije CD-jev;

– dovoljenje za uporabo slovarja je omejeno na osebo, ne več na računalnik; če ima uporabnik več naprav, naenkrat pa uporablja le eno (npr. doma, v službi, prenosni računalnik, mobilni telefon), ima brez slabe vesti omogočen dostop do slovarja z vseh naprav. Prav tako ni težav, če mora zamenjati računalnik – če le ima kopijo spletnega potrdila;

– z vidika uporabnika je začetni strošek bistveno manjši kot pri nakupu CD-ja in je zato več potencialnih kupcev.

Slaba stran je ta, da je slovar dostopen le prek spleta, a danes imajo praktično vsa podjetja neposreden dostop do interneta, obenem pa se povečuje tudi delež prebivalstva, ki ima dostop do interneta prek ADSL ali kabelske televizije, zato bo tak način dostopa vedno bolj zanimiv.

Druga slaba stran je ta, da uporabniki še niso navajeni uporabljati spletnih potrdil (težave nastajajo pri prevzemu potrdila, ne naredijo varnostne kopije, pozabijo geslo). Ker pa vedno več aplikacij zahteva spletna potrdila, bo sčasoma teh težav vedno manj.

3 Praktičen primer: *Evroterm*

Terminološka zbirka, ki uporablja izboljšave, omenjene v drugem poglavju (razen omejitve dostopa – trenutno so podatki prosto dostopni), je **Evroterm** (**http://www.gov.si/evroterm**) v kombinaciji z *Evrokorpusom* (http://www.gov.si/evrokor). *Evroterm* smo na *Centru Vlade RS za informatiko* začeli razvijati leta 2000 (Krstič 2000) v sodelovanju s službo za prevajanje, redakcijo in terminologijo pri *Službi Vlade RS za evropske zadeve*. Terminološka zbirka in korpus sta nastajala ob prevajanju pravnih aktov ES v slovenščino. Zbirka ima pretežno angleške in slovenske izraze (okoli 70.000), poleg tega je nekaj več kot 10.000 izrazov tudi v francoščini in nemščini, manjše število izrazov pa je še v drugih osmih jezikih (danščina, finščina, italijanščina, latinščina, nizozemščina, portugalščina, španščina in švedščina). Iskanje ni omejeno na iztočnice, temveč lahko iščemo po vseh sinonimih: v času pisanja tega članka baza vsebuje 69.231 vpisov (konceptov), ki so poimenovani s 76.067 slovenskimi in 72.919 angleškimi izrazi. V korpusu je okoli 22 milijonov besed. Ob iskanju posamezne besede dobimo vse možne prevode in druge podatke, ki so jih o posameznem pojmu navedli sestavljavci zbirke. Korpus je vzporedni dvojezični (angleško-slovenski), poravnava pa je na ravni prevodne enote v Tradosovem programu Translator's Workbench (ponavadi je to stavek, lahko je tudi vrstica pri naštevanju po točkah, naslov, napis pod sliko ali tabelo, vsebina celice v tabeli itd.).

3.1 Iskanje

Sodobni programi imajo velikokrat množico funkcij (ki jih uporabnik nikoli ne rabi), zato so nekatere pogosto uporabljane funkcije skrite globoko v sistemu menijev. Menim, da je pomemben napredek v nasprotno smer naredil iskalnik Google: razen nekaj vrstic besedila nad in pod iskalnim okencem je ekran praktično prazen. Po drugi strani pa strokovnjaki za terminologijo ali korpuse rabijo dodatne možnosti, s katerimi lahko filtrirajo preveliko količino izpisa, do katere pride pri preprostem iskanju in ob veliki količini podatkov v bazi. V terminološki zbirki in korpusu zato lahko uporabljamo preprosto ali izpopolnjeno iskanje.

3.1.1 Iskanje po terminološki zbirki

Preprosto iskanje: v iskalno okence vpišemo iskani izraz (besedo, del besede ali niz besed) in kliknemo iskalni gumb. Kot rezultat dobimo seznam zadetkov v vseh jezikih. Če iskane besede ni, nas program na to opozori, obenem pa poišče podobne besede in če obstajajo, jih izpiše. Ob kliku na zadetek dobimo prevod te besede in dodatne informacije o njej. V dodatnih informacijah so prevodi izbrane besede za štiri najzanimivejše jezike napisani v štirih različnih barvah, ker domnevamo, da prevajalci najbolj rabijo te podatke, v nizu besed pa obarvano besedo hitreje opazimo. Besede v drugih osmih jezikih so napisane le krepko. Besede v angleščini in slovenščini so podčrtane, kar pomeni, da jih je možno klikniti – ob tem dobimo seznam zadetkov iz korpusa, ki vsebujejo iskano besedo. Če se v polju *TermRef* pojavi veljavna oznaka predpisa po zapisu Celex, dobimo tudi povezavo na ta predpis in s klikom lahko vidimo celoten dokument.

Če po *Evrotermu* iščemo z mobilnim telefonom, je izpis seveda omejen na najnujnejše podatke: področje in izraze v vseh jezikih.

Pri razširjenem iskanju imamo več možnosti:

– določimo jezik izvirnika,

– določimo enega ali več jezikov prevoda,

– izberemo eno ali več področij,

– izberemo način ujemanja (popolno ujemanje, začetek izraza, konec izraza, del izraza ali iskanje podobnih besed),

– izberemo način izpisa:

  – seznam zadetkov z njihovimi prevodi in področji,

  – popoln izpis.

Če pri prvem načinu izpisa kliknemo označeno besedo, dobimo dodatne informacije o tej besedi (kot pri preprostem iskanju).

Če program ne najde iskane (ali njej podobne) besede v terminološki zbirki, preveri, če obstaja v korpusu.

3.1.2 Iskanje po korpusu

Pri preprostem iskanju v iskalno okence vpišemo iskani izraz (besedo, del besede ali niz besed), program pa preišče slovenski in angleški del korpusa in izpiše zadetke. Če izberemo izpopolnjeno iskanje, lahko iskanje omejimo na jezik, področje, kakovost prevoda, oznako predpisa ter izberemo eno- ali dvojezičen izpis ali le število zadetkov. Program najprej preveri, če je iskana beseda v glosarju – če obstaja, izpiše njen prevod, ki je obenem kazalec na dodatne informacije (te so enake, kot če iščemo v *Evrotermu*), nato pa se izpišejo zadetki iz korpusa. Če je program našel prevod, ta podatek uporabi pri dvojezičnem izpisu in barvno označi iskani izraz in njegov prevod. Če prevoda iskane besede ni v glosarju, je pobarvana samo iskana beseda. Če je v polju "ID" navedena veljavna oznaka predpisa po strukturi Celex, je ta pretvorjena v povezavo in če jo kliknemo, vidimo celotno besedilo predpisa.

3.1.3 Prednosti programa

V programih *Evroterm* in *Evrokorpus* so uporabljene prej naštete razširitve:

– terminološka zbirka je dvojezična (deloma dvanajstjezična), iskanje pa je možno po vseh jezikih v zbirki;

– uporabnik lahko na preprost ali kompleksnejši način preiskuje terminološko bazo in korpus, informacije se mu prikazujejo postopno: najprej seznam zadetkov, potem prevod z dodatnimi informacijami, po želji še primeri rabe ali celotno besedilo predpisa, kjer se pojavi iskana beseda;

– terminološka zbirka in korpus sta povezana, kar uporabniku prinaša dodatne koristne informacije; zbirka in korpus se lahko širita neodvisno drug od drugega;

– korpus uporablja podatke iz glosarja za preglednejši izpis;

– vsebina glosarja in dodatnih informacij se osvežuje vsaj enkrat tedensko, vsebina korpusa pa vsakih nekaj mesecev.

4 Sklep

V prispevku je naštetih nekaj možnosti, kako zasnovati nove slovarje, da nas omejitve iz knjižnih izdaj ne bodo omejevale tudi pri elektronskih slovarjih:

– iskanje po obeh jezikih v enem slovarju (in možnost iskanja po celotnem besedilu);

– iskanje podobnih besed;

– razdelitev slovarja na tri dele: glosar, dodatne informacije in primeri rabe;

– samostojen razvoj teh treh delov;

– uporaba korpusa za iskanje primerov rabe;

– korpus kot dopolnilo podatkov v slovarju in glosar kot dopolnilo podatkov v korpusu;

– redno osveževanje podatkov;

– zaščita dostopa do slovarja prek interneta s spletnimi potrdili.

Prikazano je, kako so navedeni predlogi za izboljšave uporabljeni v terminološki zbirki *Evroterm*.

Literatura

Anžlovar, Petra, 2004: Slovarji: težave se začnejo že z izborom besed. *Nedelo*, 6. junij. 27.

Brookes, Ian, 2004: Painting the Fort Bridge: Coping with Obsolescence in a Monolingual English Dictionary. *Proceedings of the Eleventh Euralex International Congress.* Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 221–231.

Drstvenšek, Nina, 2003: Vloga besedilnega korpusa pri postavitvi geselskega članka v enojezičnem slovarju. *Jezik in slovstvo* 48/5. 65–81.

Grad, Anton, Škerlj, Ružena, Vitorovič, Nada, 1997: *Veliki angleško-slovenski slovar.* Ljubljana: DZS.

Hirci, Nataša, 1999: Pogled v prihodnost: vloga prevodoslovnih besedilnih korpusov v Sloveniji. *Uporabno jezikoslovje* 7–8. Ur. I. Kovačič, I. Štrukelj. 137–154. Ljubljana: Društvo za uporabno jezikoslovje.

ISO 12616. Translation-oriented terminography, 2002. Geneva: ISO.

Jakopin, Primož, Lönneker, Birte, 2004: Query-driven Dictionary Enhancement. *Proceedings of the Eleventh Euralex International Congress*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 273–284.

Krstič, Adriana, Željko, Miran, 2000: Evroterm – terminologija EU na internetu. *Zbornik referatov posvetovanja INDO 2000*. 112–116.

Lönneker, Birte, Rozman, Katarina, 2004: Online SLO-DE-SLO: spletni slovensko-nemški in nemško-slovenski slovar. *Zbornik 7. mednarodne multikonference Informacijska družba IS 2004, zvezek B: Jezikovne tehnologije*. Ur. T. Erjavec, J. Gros. 56–63.

Vintar, Špela, 2003: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Viri na spletu

Amebis: http://www.amebis.si.

Center Vlade RS za informatiko: http://www.gov.si/cvi.

Demingov krog stalnih izboljšav: http://academic.emporia.edu/smithwil/00sum476/citeampr.htm.

Evrokorpus: http://www.gov.si/evrokor.

Evroterm: http://www.gov.si/evroterm.

Korpus *Nova beseda*: http://bos.zrc-sazu.si/s_beseda.html.

Korpusi ELAN, SVEZ-IJS in TRANS: http://nl2.ijs.si/index-bi.html.

SIGEN-CA: http://www.sigen-ca.si.

Slovar slovenskega knjižnega jezika: http://bos.zrc-sazu.si/sskj.html.

Služba Vlade RS za evropske zadeve: http://www.gov.si/svez.

Trados: http://www.trados.com.