Povezava večjezične terminološke zbirke z večjezičnim korpusom

Miran Željko

Sektor za prevajanje, Generalni sekretariat Vlade RS, Slovenija,
miran.zeljko@gmail.com

Podatki, ki jih je objavil Generalni direktorat za prevajanje v EU, so omogočili izdelavo 22-jezičnega vzporednega korpusa. V kombinaciji z večjezično terminološko zbirko je to zelo uporaben pripomoček za prevajalce. Program je prosto dostopen na spletu na naslovu http://evrokorpus.gov.si/k2.

The data published by the EU Directorate General for Translation enabled the creation of a 22-language parallel corpus. In combination with a multilingual terminology database this becomes a very useful tool for translators. The program is available online at http://evrokorpus.gov.si/k2/index.php?jezik=angl.

Ključne besede: večjezični korpus, terminološka zbirka, prevajanje
Key words: multilingual corpus, terminology database, translation

I Uvod

Slovaropisci danes slovarje pogosto naredijo na osnovi korpusov: korpus je vir besedišča za slovar, slovarska gesla so izbrana glede na pogostost pojavljanja v korpusu, primeri v slovarju so vzeti iz korpusa ipd. Med nastajanjem slovarja sta korpus in slovar neločljivo povezana – in smiselno bi bilo, da bi ostala povezana še naprej.

Žal avtorji slovarja v obdobju, ko slovar predelujejo v elektronsko obliko, ki je namenjena prodaji, slovar od korpusa odlepijo – kar je velika škoda, saj bi uporabniki slovarja zagotovo brskali po korpusu z enakim veseljem, kot so to počeli avtorji slovarja med sestavljanjem le-tega.

Idejo o povezavi slovarja in korpusa je zapisal npr. Wolfgang Teubert (Teubert 2005: 133). Eden od prosto dostopnih enojezičnih slovarjev, ki je ostal povezan s korpusom tudi za uporabnike, je Digitalni slovar nemščine (DWDS 2008).

Pri dvo- ali večjezičnih slovarjih je povezava slovarja in korpusa bolj zapletena kot pri enojezičnih, zato si za začetek lahko zadevo poenostavimo tako, da namesto s slovarjem delamo s terminološko zbirko.

II Zasnova

Ena od obsežnejših terminoloških zbirk za slovenščino je Evroterm (Evroterm 2008). Zbirka je začela nastajati med prevajanjem predpisov Evropskih skupnosti v slovenščino (Krstič 2000: 112), od leta 2004 dalje pa se dopolnjuje s terminologijo, ki se pojavlja v prevodih državnih ustanov, poleg tega pa v Evroterm vključujemo zanimive manjše terminolo-

ške zbirke in slovarje (npr. Angleško-slovenski vojaški slovar, glosar Natove terminologije, tezaver Gemet, terminologijo iz terminološke zbirke LexAlp itd.).

III Zaplet

Kmalu po nastanku Evroterma se je pokazalo, da bi bil dvojezični korpus prevodov (ki je takrat nastajal, ker so prevajalci ob prevajanju uporabljali Tradosov program Translator's Workbench) zelo uporabno dopolnilo terminološke zbirke – in tako je nastal Evrokorpus (Evrokorpus 2008). V začetku sta bila to dva neodvisna izdelka, ker pa prevajalci v obeh iščejo termine, ju je smiselno povezati. Začetki so bili precej skromni: leta 2002 (ko sta bila oba programa prvič povezana) je bila terminološka zbirka štirijezična (vsebovala je angleščino, slovenščino, nemščino in francoščino), korpus pa je bil angleško-slovenski (Željko 2002). Postopno se je terminološka zbirka razširila na 15 jezikov, v korpusni zbirki pa je do leta 2008 nastalo pet dvojezičnih korpusov (angleško-slovenski, francosko-slovenski, nemško-slovenski, italijansko-slovenski in špansko-slovenski).

Prava analogija večjezični terminološki zbirki bi bil seveda večjezični korpus – nekaj takega, kot je prevajalcem v ustanovah EU dostopno s programom Euramis (Euramis 2007).

IV Vrh

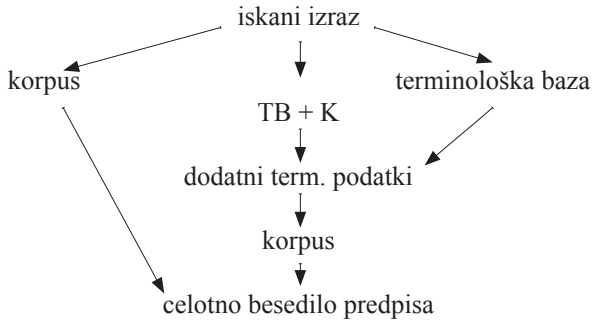
Nekaj podobnega konkordančnemu delu Euramisa je bilo možno narediti leta 2008, ko je Direktorat za prevajanje Evropske unije (DGT EU) dal v javno uporabo del podatkov iz Euramisa (Obvestilo 2008). Podatki so zapisani v večjezičnem tekstnem zapisu TMX v 22 uradnih jezikih EU in priložen je program, s katerim lahko naredimo dvojezični pomnilnik prevodov v zapisu TMX. To datoteko lahko uporabimo za polnjenje pomnilnika prevodov za tisti program, ki ga uporabljamo sami. Sparimo lahko poljubna jezika, tudi taka, ki sicer še nista bila v stiku (npr. slovenščina – estonščina).

Zapis je preprost, zato je podatke razmeroma preprosto uporabiti za izdelavo večjezičnega korpusa. Zaradi precejšnje količine podatkov obdelava sicer traja nekaj časa, na srečo pa so ustvarjalci to predvideli in so podatke razdelili na 12 datotek. V različnih jezikih je različno število segmentov – ker je bila največkrat izvirni jezik angleščina, je v njej največ podatkov. Zanimivo je, da je v jezikih novih članic EU (vključenih leta 2004 in kasneje) več podatkov kot v jezikih mnogih starih članic – ker so nove članice že uporabljale programe za delo s pomnilniki prevodov, ko so prevajalci iz »starih« jezikov še vedno delali ročno.

V Razplet

Po izdelavi druge verzije programov Evroterm in Evrokorpus (leta 2004) se je pokazalo, da bi bilo možno narediti en uporabniški vmesnik za iskanje po terminološki zbirki in po korpusu – zato sem to priložnost izkoristil še za to, da ima nova verzija programa samo en uporabniški vmesnik. Program je dostopen na naslovu http://evrokorpus.gov.si/k2. Uporabniku omogoča tri načine iskanja:

1. iskanje samo v terminološki zbirki,
2. iskanje samo v korpusu,
3. iskanje v terminološki zbirki in v korpusu.



Slika 1: Načini iskanja po večjezičnem korpusu

Pri vnosu podatkov mora uporabnik izbrati jezik izvirnika in poljubno število jezikov prevoda. Po želji lahko iskanje omeji na kakšno oznako Celex (ali na del take oznake). Omejitev obsega izpisa in izpis celotnega števila zadetkov sta namenjena samo temu, da skrajšata čas iskanja, kadar uporabnik ne potrebuje obsežnega izpisa. Pri korpusu lahko izberemo eno- ali večjezični izpis.

Pri prvem načinu iskanja (iskanje v terminološki zbirki) dobimo v izpisu vse tiste termine (in njihove prevode), ki ustrezajo iskanemu pogoju. Po želji pa s klikom dobimo iz Evroterma tudi druge podatke o iskanem terminu (npr. definicijo, zanesljivost, referenčni dokument, področje uporabe ipd.). V terminološki zbirki je bistveno manj podatkov kot v korpusu in terminološki podatki so dobro urejeni, zato je ta način iskanja najhitrejši.

Pri iskanju samo v korpusu nam program izpiše korpusne rezultate, ki ustrezajo iskanemu pogoju. Iskana beseda v jeziku izvirnika je obarvana modro, da jo hitreje najdemo. Kadar zahtevamo večjezični izpis, program izpiše tiste segmente, pri katerih obstaja prevod vsaj v enem jeziku. Če je skupno število jezikov do pet, so rezultati izpisani vzporedno (drug poleg drugega), ker je tak izpis preglednejši (ob predpostavki, da je vrstni red besed v različnih jezikih podoben). Če je jezikov več, so segmenti zapisani drug pod drugim. Ob izpisu dobimo prek oznake Celex povezavo na celotno besedilo predpisa, iz katerega izhaja dotični segment. Za bazo celotnih besedil predpisov uporabljamo Eur-Lex (Eur-Lex 2008). Izpis je v začetku enojezičen, lahko pa si naredimo dvojezičnega, če na začetku besedila predpisa izberemo poljubna jezika.

Pri enojezičnem izpisu dobimo centriran izpis KWIC (Key Word In Context); levo in desno od iskane besede je do 40 znakov (če želimo videti celotne segmente, izberemo dvojezični izpis in za jezik prevoda izberemo isti jezik kot za izvirnik).

Iskanje v terminološki zbirki in korpusu združuje oba prej opisana izpisa. Dodatna vrednost tega izpisa je ta, da je prevod iskane besede v prevedenem segmentu obarvan modro (če ga program najde).

Korpus sicer ni lematiziran, program za iskanje po njem pa pozna nekaj pravil, s katerimi najde tudi nekatere sklanjane samostalnike in pridevnike za slovenščino (npr. *država – države – državo* itd., *kovanec – kovanca* itd., *Evropska unija – Evropske unije* itd.), angleščino (*company – companies* ipd.) in nemščino (*Europäische Union – Europäischen Union* ipd.). Zaradi večje verjetnosti napak je ta logika izključena pri kratkih besedah (npr. beseda *act* se sicer pojavlja pogosto, vendar se v angleščini na *-act* konča veliko besed), kljub temu pa se

včasih pojavi napaka tudi pri daljših besedah (npr. iščemo *kovanec*, pa dobimo med zadetki tudi *preiskovanec*). Poleg tega se včasih zgodi, da so pri različnih besedah nekatere oblike enake (npr. *porok* je kot samostalnik moškega spola v imenovalniku enak samostalniku ženskega spola *poroka* v rodilniku množine) – tudi v takih primerih program optimistično izpiše vse, kar najde.

VI Razsnova

Prednosti uporabe večjezičnega vzporednega korpusa za prevajalce v primerjavi z dvojezičnim vzporednim korpusom so:

1. Program omogoča drugačen način prevajanja kot doslej. Že prej je sicer veljalo, da je za prevajalca koristno, če obvlada več tujih jezikov – z uporabo tega korpusa pa se pokaže, da je znanje več tujih jezikov koristno, tudi če prevajalec prevaja samo iz enega jezika, saj si pri dvoumnih besednih zvezah v izvirniku lahko pomaga s prevodi v drugih jezikih. Prav tako mu drugi tuji jeziki pomagajo, če dvomi o pravilnosti prevoda iz korpusa ali pa kadar v ciljnem jeziku sploh ni prevoda. Res je, da si v takih primerih lahko pomaga z drugimi pripomočki, npr. pri prevodu iz angleščine v slovenščino tudi z angleško-nemškimi slovarji in angleško-nemškimi korpusi – je pa delo gotovo hitrejše, če je vse skupaj na enem mestu.
2. Za nekatere jezikovne kombinacije Slovenci sploh nimamo slovarjev (npr. baltski jeziki, malteščina), zato je tak način iskanja ustreznic pravzaprav edina možnost.
3. Programi za delo s pomnilniki prevodov nam sicer omogočajo konkordančno iskanje po pomnilniku in to je praviloma hitrejše kot iskanje po korpusu prek interneta – vendar so tovrstni pomnilniki praviloma dvojezični (pri čemer npr. v SDL Tradosu rabimo celo dva pomnilnika(!), če hočemo npr. v angleško-slovenskem pomnilniku poleg angleških besed iskati tudi slovenske). Dvojezični korpus je morda še neke vrste (slaba) konkurenca dvojezičnemu pomnilniku prevodov, 22-jezični vzporedni korpus pa med obstoječimi najbolj razširjenimi pomnilniki prevodov nima prave konkurence.

Gotovo ima 22-jezični korpus v obstoječi obliki tudi pomanjkljivosti:

1. V korpusu so napake – ta očitek seveda velja za vse korpuse, ne le za tega. Zanimivo je, da so napake celo v kratkem besedilu, s katerim je DGT EU želel pokazati, kako je vzorčen stavek napisan v 22 različnih jezikih (Referenčni stavek 2008). Glede na doslej najdene napake ocenjujem, da (vsi) prevodi niso bili dodatno pregledani (strokovno, pravno ali jezikovno). Upoštevati je treba tudi, da je bila večina prevodov narejena iz angleščine, kar pomeni, da pri sparjenju stavka v dveh ali več jezikih (med katerimi ni angleščine) lahko pride do razlik, ki so razumljive le, če upoštevamo postopek dela. Vse to pomeni, da je pri uporabi vsakega izraza potrebna neka količina skepse, posebej če prevoda iskanega izraza ne poznamo – v takih primerih je vedno koristno, da ga preverimo še kje druge, če je to le možno. To je še pomembneje, če si prevajalec iz prosto dostopnih podatkov naredi pomnilnik prevodov – obvezno je treba preveriti tudi 100-odstotne zadetke.
2. Hitrost – povečati bo treba hitrost iskanja.

3. Korpus ni lematiziran, zato uporabnik dobi manj zadetkov, kot bi jih lahko.
4. V dvojezičnih korpusih je več besed, kot jih je (za isti jezikovni par) v 22-jezičnem korpusu – v prihodnosti bo treba vse posamezne dvojezične korpuse združiti v enega večjezičnega, pri čemer je pomembno, da izločimo duplikate.

Možnosti za izboljšave v prihodnosti je torej dovolj.

VII Sklep

Za prevajalce so poleg slovarjev, glosarjev, terminoloških in podobnih besednih zbirk zelo pomembni tudi korpusi. Praviloma korpusi obstajajo ločeno od besednih zbirk – so pa precej bolj uporabni, če sta ti dve vrsti zbirk medsebojno povezani. Za prevajalce so bolj kot enojezični korpusi uporabni dvojezični, ker iz njih vidijo rabo besede ali besedne zveze v obeh jezikih v kontekstu in so zato naravno dopolnilo slovarjev. Korak naprej so večjezični korpusi, ker prevajalcem besedilo v drugih jezikih pomaga pri boljšem razumevanju izvirnika. Prevajalci v EU imajo v ta namen program Euramis, ki pa ni dostopen javnosti. Med prosto dostopnimi izdelki je eden prvih korakov v tej smeri večjezični korpus na naslovu http://evrokorpus.gov.si/k2.

VIII Literatura

Digitalni slovar nemščine: http://www.dwds.de

Euramis: http://europa-eu-audience.typepad.com/en/files/euramis_Presentation_2007_03.ppt

Eur-Lex: http://eur-lex.europa.eu

Evrokorpus: http://evrokorpus.gov.si

Evroterm: http://evroterm.gov.si

Adriana KRSTIČ, Miran ŽELJKO, 2000: Evroterm – terminologija EU na internetu. *Zbornik referatov s posvetovanja INDO 2000*. 112–116.

Obvestilo DGT EU o javnem dostopu do prevodov predpisov EU v vseh uradnih jezikih EU: http://europa.eu/rapid/pressReleasesAction.do?reference=IP/08/60&format=HTML&language=SL

Referenčni stavek, napisan v več jezikih: http://langtech.jrc.it/Documents/0801_DGT-TM_Sample-translation-unit_all-languages.doc

Wolfgang TEUBERT, 2005: Korpusno jezikoslovje in leksikografija. Vojko Gorjanc, Simon Krek (ur.): *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina. 103–136.

Miran ŽELJKO, 2002: Pripomočki na spletu za prevajalce zakonodaje EU. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Jezikovne tehnologije: zbornik konference*. Ljubljana: Institut Jožef Stefan.