FILOLOGIJA 38—39, Zagreb 2002

UDK 81'374.82
Pregledni članak
Primljen 31.I.2002.
Prihvaćen za tisak 20.V.2002.

Ivana Simeon
Zavod za lingvistiku
Filozofski fakultet Sveučilišta u Zagrebu
Ivana Lučića 3, HR-10000 Zagreb

PARALELNI KORPUSI I VIŠEJEZIČNI RJEČNICI

Paralelni korpus jest dvojezični ili višejezični korpus koji sadrži niz tekstova na dva ili više jezika.
Nakon što se prikupe i obrade (što uključuje uklanjanje pogrešaka, segmentaciju, sravnjivanje međusobno podudarnih segmenata i anotaciju), paralelni korpusi predstavljaju važan alat za istraživanje terminologije, kontrastivnu lingvističku analizu, definiranje prijevodnih ekvivalenata, te su stoga od neprocjenjive vrijednosti pri sastavljanju dvojezičnih i višejezičnih rječnika.
Ovaj rad daje pregled metoda za pripremu i obradu paralelnih korpusa te za njihovu uporabu u višejezičnoj leksikografiji.

1. Uvod

Posljednjih dvadesetak godina, u leksikografskoj metodologiji sve se više primjenjuje korpusni pristup, o čemu svjedoče brojni rječnici koji svoju utemeljenost na korpusu ističu kao glavnu prednost pred tradicionalno (uglavnom intuitivno) sastavljenim rječnicima. Leksikografi se korpusnim pristupom služe posebno pri sastavljanju rječnika namijenjenih onima koji uče neki strani jezik, te definicije pojmova potkrepljuju ne samo ovjerenim već i potvrđenim, realnim primjerima uporabe.

Dok je pri sastavljanju jednojezičnih rječnika glavni resurs jednojezičan korpus, u dvojezičnoj i višejezičnoj leksikografiji primjenjuju se paralelni korpusi, odnosno korpusi jezika koji su uključeni u neki rječnik. U ovom članku razjašnjavaju se osnovni pojmovi korpusne metodologije, daje se pregled metoda za obradu paralelnih korpusa i obrazlažu prednosti korpusnog pristupa leksikografiji.

2. Paralelni korpusi

Paralelan korpus jest dvojezični ili višejezični korpus koji sadrži niz tekstova pisanih na dva ili više jezika. Postoji nekoliko osnovnih tipova takva korpusa:

- paralelni korpus koji sadrži tekstove izvorno napisane na jeziku A i njihove prijevode na jezik B (te C, D, ...),
- paralelni korpus koji sadrži jednaku količinu tekstova izvorno napisanih na jezicima A i B te njihove prijevode,
- paralelni korpus koji sadrži samo prijevode na jezike A, B i C, dok je tekst bio izvorno napisan na jeziku Z[1],

Paralelni korpusi predstavljaju bogat lingvistički resurs, jer sadrže opsežnu količinu podataka o stvarnoj jezičnoj uporabi. Navest ću neka od brojnih područja njihove primjene: razvoj sustava za strojno i strojno potpomognuto prevođenje, kontrastivna i terminološka istraživanja, glotodidaktika te **dvojezična i višejezična leksikografija**.

3. Obrada paralenih korpusa

Paralelnokorpusna građa zahtijeva nekoliko tipova obrade kako bi iz nje dobiveni podaci bili pouzdani i iskoristivi u leksikografskom radu. Primarna obrada – predobrada – ima nekoliko koraka; to su

- prikupljanje samog korpusa odnosno usporednih tekstova,
- uklanjanje pogrešaka i informacija o formatiranju iz tekstova,
- segmentacija teksta i obilježavanje na razini odlomaka i rečenica,
- opojavničenje (*tokenization*), odnosno segmentiranje teksta na pojavnice,
- sravnjivanje (*alignment*), odnosno povezivanje segmenata polaznog teksta s odgovarajućim segmentima ciljnog teksta ili tekstova.

3.1. Sravnjivanje

Sravnjivanje je postupak kojime se segmenti (odlomci, rečenice i eventualno riječi) u polaznom tekstu povezuju s istorazinskim i odgovarajućim segmentima u ciljnom tekstu. Najčešći i najuspješniji oblik sravnjivanja jest na razini rečenica. Sravnjivanje može biti potpuno automatizirano, iako je najčešće poluautomatizirano, odnosno potrebna je ljudska intervencija, obično u završnoj fazi.

Činjenica koja otežava potpunu automatizaciju sravnjivanja jest da rečenice nisu uvijek prevedene po načelu 1:1. Prevoditelji se često odlučuju za različite izmjene kao što je dokidanje (1:0), umetanje (0:1), razbijanje jedne rečenice u dvije ili više rečenica (1:x), spajanje rečenica (x:1) ili prestrukturiranje rečenica (x:y). Program za sravnjivanje (koji, barem zasad, nema znanje o jeziku i izvanjezičnom univerzumu s pomoću kojeg bi mogao razriješiti takve poteškoće) može pogriješiti pri utvrđivanju odgovarajućih segmenata. Kako bi

[1] Teubert 1996:238–265.

se izbjegla mogućnost lančanog nizanja pogrešaka, odlomci se definiraju kao čvrsta referentna točka (*hard link*), a rečenice kao promjenjiva referentna točka (*soft link*). Tako se eventualne pogreške zadržavaju unutar jednog odlomka, a ne prenose se u sljedeći.

3.1.1. Metode sravnjivanja

Metode sravnjivanja ugrubo se mogu podijeliti na metode koje se oslanjaju na srodnost i/ili sličnost riječi u jezicima koji sačinjavaju paralelni korpus i na jezično neovisne metode. U prvu skupinu spadaju:

- metoda srodnih i *sidrišnih* riječi, koja je svoju primjenu našla u programu *Char_Align* tvrtke AT&T i koja uspoređuje tekstove pisane na srodnim jezicima na razini pismena, tražeći srodne riječi koje zatim postaju *sidrišta* prema kojima se vrši sravnjivanje;
- bitekstualno preslikavanje (*Smooth Injective Map Recognizer* – SIMR) – primjenjuje se kod dvojezičnih tekstova koji predstavljaju dvodimenzionalan *bitekstualan prostor* kojemu su osi položaji pismena u polaznom i u ciljnom jeziku; algoritmom se pretražuju stvarne točke podudarnosti.

Jezično neovisne metode imaju tu prednost da su primjenjive na bilo koji jezik, odnosno bilo koju kombinaciju jezika. Najzastupljenije su ove dvije metode:

- *Church–Galeov algoritam*[2], koji se temelji na pretpostavci da odgovarajući segmenti paralelnih tekstova imaju sličnu duljinu u pismenima. Taj je algoritam poslužio kao temelj za sravnjivački program *Vanilla Aligner*[3] koji su razvili Daniel Ridings i Pernilla Danielsson sa sveučilišta u Göteborgu, a koji je vrlo uspješno primijenjen za sravnjivanje hrvatsko-engleskog i hrvatsko-slovenskog paralelnog korpusa u Zavodu za lingvistiku Filozofskoga fakulteta u Zagrebu.
- Metode *K-Vec* i *DK-Vec*, koje je razvila Pascale Fung sa sveučilišta Columbia u New Yorku. Obje se metode temelje na statističkom pristupu i njima se pretražuju sličnosti u distribuciji riječi u polaznom i u ciljnom tekstu, pri čemu je metoda DK-Vec pogodna ne samo za izravno paralelne, već i za ugrubo podudarne korpuse[4].

3.1.2. Pohranjivanje i pristup podacima

Za kodiranje podataka sadržanih u paralelnim korpusima u posljednje se vrijeme sve više koristi *Extensible Markup Language* (XML). Prednost tog jezika

[2] Church–Gale 1993.
[3] Ridings–Danielsson 1999.
[4] Tiedemann 1998.

za obilježavanje jest to što on čuva informacije o hijerarhijskoj strukturi teksta, fleksibilan je te olakšava upravljanje podacima.

Osnovni problem koji se javlja pri uporabi paralelnih korpusa jest njihova obimnost, što usporava pristup i obradu. Taj se problem može riješiti tako da se tekstovni podaci pohrane na jednome mjestu, a strukturni na drugomu. Pri tome se odabranim segmentima dodjeljuje jedinstvena identifikacijska oznaka koja povezuje taj segment sa strukturalnom informacijom. Rezultat takva pristupa jest brz pristup i brza pretraživost paralelnih korpusa.

4. Primjena paralelnih korpusa u višejezičnoj leksikografiji

Paralelni korpusi omogućavaju leksikografima da proučavaju riječi i kombinacije riječi te njihove prijevodne ekvivalente unutar konteksta u kojem se stvarno pojavljuju. Na taj način oni korisniku mogu ponuditi važne podatke o nekim aspektima značenja riječi koji bi im inače možda promakli te se usredotočiti na značenja i konstrukcije čija je uporaba najrasprostranjenija.

Nadalje, moguće je automatski ili poluautomatski generirati rječničke natuknice iz paralelnih korpusa, o čemu će biti više riječi u nastavku.

Naposljetku, paralelni se korpusi uspješno koriste u izgradnji dinamičkih *on-line* rječnika.

5. Izdvajanje prijevodnih ekvivalenata

Sravnjivanje paralelnih tekstova i izdvajanje prijevodnih ekvivalenata iz tih tekstova međusobno su komplementarni. Naime, za uspješno izdvajanje prijevodnih ekvivalenata nužno je da tekstovi budu kvalitetno i precizno sravnjeni. S druge strane, ako je već izdvojen skup prijevodnih ekvivalenata, to uvelike olakšava postupak sravnjivanja, jer ti ekvivalenti predstavljaju referentne točke u tekstovima.

Tri su osnovna pristupa izdvajanju prijevodnih ekvivalenata:

- izdvajanje putem uzastopnog smanjenja veličine,
- izdvajanje na temelju srodnosti jezikā,
- statistički pristup[5].

Izdvajanje metodom uzastopnog smanjenja veličine pogodno je za visokostrukturirane tekstove, kao što je tehnička dokumentacija i sl. U prvoj fazi izdvajaju se parovi sravnjenih pojavnica tipa 1:1, 1:x i x:1 te se prikuplja osnovni skup, odnosno rječnik prijevodnih ekvivalenata. S pomoću tog se rječnika u drugoj fazi, ponavljanjem postupka, analiziraju preostale sravnjene pojavnice kako bi se iz kompletnog skupa uklonile one koje su već uvrštene u temeljni rječnik. Zatim se dobiveni ekvivalenti tipa 1:1 izdvajaju i pridodaju

[5] Tiedemann 1998.

skupu poznatih prijevodnih jedinica. Taj se prošireni skup opet koristi za analizu preostalih potencijalnih ekvivalenata. Postupak se ponavlja sve dok se ne isključe svi parovi tipa 1:1. U primjeni na korpus Scania[6] prikupljen na sveučilištu u Uppsali, taj je algoritam postigao visoku preciznost.

Izdvajanje na temelju srodnosti jezikā prikladno je samo za genetski bliske jezike, a najbolje rezultate postiže kod stručnih tekstova zbog internacionalizacije terminologije. Taj su pristup primijenili I. Dagan i K. Church iz AT&T Laboratories pri izradi sustava Termight[7], koji se temelji na programu Char_Align; tim se sustavom uspoređuju srodni nizovi pismena i stvara popis prijevodnih kandidata koji se rangiraju prema frekvenciji.

Leksikograf se može odlučiti za jednu od statističkih metoda izdvajanja prijevodnih ekvivalenata koje su, kao što je već istaknuto, neovisne o jeziku, odnosno primjenjive na bilo koju kombinaciju jezika.

Navest ću dva algoritma za izdvajanje prijevodnih ekvivalenata. Prvi je od njih Diceov koeficijent, kojime se mjeri zajedničko pojavljivanje parova riječi ili parova skupina riječi u podudarnim segmentima teksta.

(1) Diceov koeficijent[8]

$$Dice(x,y) = \frac{2P(x,y)}{P(x)+P(y)}$$

gdje je $P(x,y)$ vjerojatnost zajedničkog pojavljivanja x i y u podudarnim segmentima, dok su $P(x)$ i $P(y)$ pojedinačne vjerojatnosti pojavljivanja x i y, pri čemu x i y mogu biti i riječi i skupine riječi.

Drugi algoritam za izdvajanje prijevodnih ekvivalenata jest uzajamna obavijesnost (*Mutual Information* — MI).

(2) Uzajamna obavijesnost[9]

Vrijednost uzajamne obavijesnosti pokazuje u kolikom se broju slučajeva dvije pojavnice — kandidati za prijevodne ekvivalente — pojavljuju zajedno. Dakako, niskofrekventne pojavnice izbacuju se iz analize, jer njihov visoki rezultat nije statistički značajan.

6. Otkrivanje i uklanjanje pogrešaka

Pri izdvajanju prijevodnih ekvivalenata mogu se javiti pogreške, odnosno pogrešno spareni ekvivalenti. Kako bi se te pogreške eliminirale, primjenjuje se nekoliko filtara koji pročišćuju rezultate:

[6] Scania 2001.
[7] Dagan–Church 1994.
[8] Smadja 1996.
[9] Church–Hanks 1989.

- Filtar temeljen na duljini: omjer razlike u duljini računa se dijeljenjem duljine kraćeg niza s duljinom duljeg niza. Par koji ima najveći rezultat najvjerojatnije predstavlja najbolji prijevodni ekvivalent.
- Filtar sličnosti: na izdvojene parove primjenjuju se algoritmi za uspoređivanje riječi. Najviši rezultat ukazuje na najvjerojatniji prijevodni ekvivalent.
- Filtar temeljen na frekvenciji: izračunava se apsolutna frekvencija i frekvencija zajedničkog pojavljivanja za izdvojene parove. S tim vrijednostima izračunava se Diceov koeficijent i uklanjaju se potencijalni prijevodni ekvivalenti s niskim rezultatom.
- Kombinirani filtar: spomenuti filtri mogu se kombinirati – jedan je pristup prikladniji za utvrđivanje najvjerojatnijeg prijevoda, a drugi za usporedbu alternativnih prijevoda s tim najvjerojatnijim prijevodom.
- Filtar podskupova: potencijalni prijevodi mogu biti nepotpuni, pa je potrebno izbaciti prijevod koji je uključen u drugi prijevod.[10]

7. Zaključne primjedbe

Uporaba korpusa donijela je višejezičnoj leksikografiji brojne pozitivne promjene. Kao prvo, paralelni korpus omogućuje trenutačan pristup velikoj količini jezičnih podataka. Nadalje, ti su podaci relevantni za stvarnu uporabu leksičkih jedinica u prijevodima i daju presjek aktualne prevodilačke prakse. Naposljetku, digitalna priroda paralelnih korpusa omogućava lakše i brže utvrđivanje prijevodnih ekvivalenata, sastavljanje elektronskih rječnika i jednostavno i brzo ažuriranje baze jezičnih podataka. Stoga su paralelnokorpusni projekti od velike važnosti za leksikografe, ali i za održavanje koraka sa sve zahtjevnijom i opsežnijom višejezičnom komunikacijom.

[10] Tiedemann 1998.

Literatura

Church, Kenneth, Patrick Hanks. 1989. Word association norms, mutual information, and lexicography, *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 76–83.

Church, Kenneth, William Gale. 1993. A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics* 19:1.

Dagan, Ido, Kenneth Church. 1994. Termight: Identifying and translating technical terminology. *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP)*, 34–40.

Danielsson, Pernilla, Daniel Ridings. 1999. *Practical Presentation of a »Vanilla« Aligner*, (11.X.1999.), *http://nl.ijs.si/telri/Vanilla/doc/ljubljana/*, siječanj 2002.

The Scania Project, 2001. *http://stp.ling.uu.se/~corpora/scania/*, siječanj 2002.

Smadja *et al.* 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics* 22:1, 3–38.

Teubert, Wolfgang. 1996. Comparable or Parallel Corpora?, *International Journal of Lexicography* 9:3, 238–265.

Tiedemann, Jörg. 1998. *Extraction of Translation Equivalents from Parallel Corpora, http://numerus.ling.uu.se/~joerg/paper/Nodalida98/Nodalida98.html*, siječanj 2002.

Parallel corpora and multilingual dictionaries

Summary

A parallel corpus is a bilingual or multilingual corpus, containing texts written in two or more languages.

After they are compiled and processed (which includes correcting errors, segmentation and alignment of corresponding segments), parallel corpora provide a valuable tool for terminological research, contrastive linguistic analysis, determining translation equivalents and are therefore an important resource for bilingual and multilingual lexicography.

This paper gives a review of methods for preparation and processing of parallel corpora, as well as their use in multilingual lexicography.

Ključne riječi: paralelni korpusi, sravnjivanje, prijevodni ekvivalenti, višejezična leksikografija, višejezični rječnici

Key words: parallel corpora, alignment, translation equivalents, multilingual lexicography, multilingual dictionaries