Univerza v Ljubljani
Naravoslovnotehniška fakulteta
Oddelek za tekstilstvo

Mirica DEBELJAK

STROJNO PREVAJANJE

seminarska naloga
(Jezikovne tehnologije)

Ljubljana, 2005

KAZALO

1	**UVOD**	1
2	**SPLOŠNI ORIS PODROČJA**	2
3	**RAČUNALNIŠKE TEHNOLOGIJE ZA PREVAJANJE IN TERMINOLOGIJO**	5
4	**STROJNO PREVAJANJE**	8
4.1	**Zapletena slovenščina**	14
4.2	**Prevajalni sistem PRESIS**	14
5	**RAČUNALNIŠKO PODPRTO PREVAJANJE**	18
5.1	**Pomnilniki prevodov**	18
5.2	**Terminološki programi**	19
6	**KORPUSI**	20
6.1	**Korpus slovenskega jezika FIDA**	25
7	**STANDARDI ZAPISA JEZIKOVNIH PODATKOV**	26
8	**ZAKLJUČEK**	29
9	**LITERATURA**	30

1. UVOD

Povsod po svetu je vse več povpraševanja po prevajalskih storitvah. Zaradi vse hitrejšega razvoja komunikacijsko-informacijskih tehnologij ter gospodarskih in političnih postopkov evropskega in svetovnega združevanja se jezikovne in računalniške tehnologije razvijajo zelo hitro.

Jezikovne tehnologije so vse tiste informacijske tehnologije, ki se ukvarjajo z naravnimi jeziki, med katerimi sodi tudi slovenščina. Jezikovne tehnologije postajajo vse pomembnejši sestavni del informacijskih tehnologij, to pa je v osnovi omogočila šele dovolj velika procesna moč sodobnih računalnikov in drugih naprav, ki so hkrati sposobne hraniti velikansko število informacij.

Na področju jezikovnih tehnologij deluje veliko število podjetij, organizacij in ustanov, razvijalci in proizvajalci se srečujejo na mnogih mednarodnih konferencah, kakovost številnih rešitev ter sistemov s tega področja je vsaj za večino glavnih svetovnih jezikov na precej visoki ravni.

Ker pa je pri nas računalniška podpora prevajalskemu in terminološkemu delu novo področje, je njeno poznavanje in raba zaenkrat še v začetni fazi. Težava, s katero se srečuje razvoj teh tehnologij pri nas, je neosveščenost. Slabo poznavanje prevajalskih orodij je posledica majhnosti trga in zamude pri vključevanju v večje evropske in svetovne združbe. Ponudnikov prevajalske in terminološke programske opreme na slovenskih tleh skoraj ni. Prevajalec lahko pride v stik s temi orodji prek študijskih programov ali na delovnem mestu.

Poleg osveščenosti pa morajo biti za uporabo računalniške podpore izpolnjeni tudi drugi pogoji. Uporabnik mora vložiti denar za nakup programskega orodja (ne vedno, saj je na spletu določena programska oprema brezplačna), čas za učenje dela s programom in za gradnjo baze podatkov, poleg tega pa mora znati izrabljati jezikovne vire in imeti določene računalniške spretnosti.

Računalniške tehnologije terminološkemu in prevajalskemu delu omogočajo kompaktno, poceni in dolgoročno shranjevanje, hitro obdelavo velikih količin podatkov, lažjo izmenjavo in distribucijo podatkov, posodabljanje, raznovrstno iskanje po podatkovnih bazah, urejanje vnosov delov besedil, povezave med njimi in z drugimi programi (npr. urejevalnikom besedil, lokalizacijskimi programi) ter avtomatizacijo iskanja terminologije in prevodnih ustreznic.

2. SPLOŠNI ORIS PODROČJA

Alan K. Melby, predsednik Odbora za prevajanje in računalnike Ameriške zveze prevajalcev (*American Translators Association Translation and Computers Committee*) in član pomembnih institucij, ki se ukvarjajo s prevajanjem in terminologijo (*Association for Computing in the Humanities*, *Association for Computational Linguistics*, *Association for Literary and Linguistic Computing*, *ISO* itn.), razdeli računalniško podporo prevajanju na **osem tipov**, ki so v uporabi na treh stopnjah prevajanja, na ravni posameznega izraza in celega segmenta (tj. del besedila, večji od izraza, ponavadi stavek):

INFRASTRUKTURA

	IZRAZ	SEGMENT
PRED PREVAJANJEM	• Pol(samodejno) luščenje terminoloških kandidatov • Raziskovanje terminologije	• Poravnava in označevanje prejšnjih izhodiščnih in ciljnih besedil ter segmentacija novega izhodiščnega besedila
MED PREVAJANJEM	• Samodejno iskanje in vstavljanje izrazov	• Pomnilnik prevodov • Strojni prevajalnik
PO PREVAJANJU	• Pregled terminološke doslednosti in nedovoljene terminologije	• Zaznava manjkajočega segmenta ter pregledi oblike in slovnice

PREGLED NAD PREVAJALSKIM POSTOPKOM IN OBRAČUN

Slika: Osem tipov prevajalskih tehnologij.

1. Infrastruktura (ni neposredno del prevajanja) je pomembna zlasti v večjezikovnih situacijah. Elementi infrastrukture morajo biti čim bolj enotni, med seboj in s prevajalskim postopkom. Ti so:

- **orodja za pripravo elektronske oblike besedila**:
  - orodja za skeniranje in pretvorbo grafičnega zapisa v elektronsko besedilo (optično čitanje znakov – OCR)

- **orodja za ustvarjanje in upravljanje dokumenta**:
  - urejevalnik besedil (npr. Microsoftov Word, Sunov StarOffice) z najpomembnejšimi funkcijami: oblikovanje dokumentov, štetje besed in znakov (za izračun prevajalskih strani), nastavitev jezikovne podpore v operacijskem sistemu, tuji nabori znakov, primerjava različic dokumentov itd.
  - jezikovna orodja, integrirana v urejevalnik besedil: črkovalnik, osebni slovarji, preverjanje slovnice, slovar sopomenk, tezaver, delilnik besed, prevajanje itd.

- **podatkovne zbirke**:
  - elektronski slovarji (eno-, dvo- in večjezični slovarji) kot npr. Slovensko-angleški, Collins Cobuild, LINA (zadnji in najnovejši program istega podjetja, ki omogoča delo z različnimi slovarji hkrati), s katerimi je mogočih več načinov iskanja, dodajanje opomb itd.
  - drugi elektronski viri na zgoščenkah: enciklopedije (Encarta, World Atlas, Britannica), serijske publikacije (Uradni list) itd.
  - splošni programi za izdelavo podatkovnih baz: Microsoftov Excel in Access, Oracle 8i itd.

- **telekomunikacije** (internet/intranet, elektronska pošta, protokol za prenos datotek (FTP), telefonski stiki), ki nam omogočajo dostop tudi do drugih prevajalskih virov in programov:
  - podatkovne zbirke na internetu: slovarji, tezavri, terminološke baze, bibliografske baze (npr. COBISS), zbirke besedil (različni korpusi), primerljiva besedila itn.
  - programi za prevajanje in urejanje terminologije na internetu: programi s pomnilniki prevodov, terminološki programi, strojni prevajalniki, orodja za lokalizacijo računalniških programov (prilagoditev programske opreme jezikovnim in kulturnim zahtevam okolja, kjer se uporablja) idr.
  - drugi viri na internetu: serijske publikacije, domače strani proizvajalcev programske opreme, prevajalskih inštitutov, društev, agencij, konferenc, kongresov, druga poročila iz akademsko-raziskovalnih krogov, informacije o prevodoslovju, jezikoslovju, jezikovnih in računalniških tehnologijah ipd.

2. Pred prevajanjem na ravni izraza: (Pol)samodejno luščenje terminoloških kandidatov in raziskovanje terminologije

S tema dvema orodjema določamo, katere prevodne ustreznice bi lahko vključili v terminološko bazo. Ko orodje za luščenje izrazov (ali kateri drug program) prepozna izhodiščni izraz, uporabimo orodje za raziskovanje terminologije, ki določi ciljnega. Program za luščenje terminoloških kandidatov je po nalogi podoben črkovalniku, vendar deluje dosti bolje, saj prevajalcu ponudi tudi izraze, iz katerih lahko nastanejo novi večbesedni izrazi (npr. ne odločamo se le med izrazoma thermal in layer, ampak lahko izberemo tudi celo besedno zvezo thermal layer). Orodje za iskanje terminologije lahko išče v več virih, npr. v že prevedenih besedilih, na internetu, v večjezikovnih besedilnih zbirkah ipd.

3. Med prevajanjem na ravni izraza: Samodejno iskanje in vstavljanje izrazov.

Ta postopek bi lahko opisali kot strojno prevajanje na ravni izrazov. Ko prevajalec začne z urejanjem in prevajanjem segmenta, se na zaslonu pojavijo ciljni izrazi. Prevajalec izbere pravega in ga samodejno prenese v dokument brez tveganja napak pri črkovanju. Takšno iskanje terminologije omogoča dosledno uporabo izrazov.

4. Po prevajanju na ravni izraza: Pregled terminološke doslednosti in nedovoljene terminologije.

Pregledovalci se sprožijo, ko je prevod že končan. Označijo terminološke nedoslednosti in izraze, ki niso primerni oz. dovoljeni v besedilu.

5. Pred prevajanjem na ravni segmenta: Poravnava in označevanje prejšnjih izhodiščnih in ciljnih besedil ter segmentacija novega izhodiščnega besedila.

Poravnava in označitev segmentov izhodiščnega in ciljnega besedila za ponovno uporabo sta nujni za pravilno delovanje programa s pomnilnikom prevodov. Označeni pari so uporabni tudi za iskanje izrazov.

6. Med prevajanjem na ravni segmenta: Pomnilnik prevodov in strojni prevajalnik

Ko so segmenti označeni in poravnani, pomnilnik prevodov pregleda prejšnja prevedena besedila, primerja nove segmente s segmenti v svoji bazi in samodejno prikliče tiste, ki niso (veliko) spremenjeni ter jih pripravi za ponovno uporabo. Pri prevedenih besedilih, ki potrebujejo le nekaj majhnih popravkov, je pomnilnik prevodov še posebej učinkovit.

Strojni prevajalnik algoritmično obdela izhodiščno besedilo, prepozna besede in razmerja med njimi, izbere izraze v ciljnem jeziku, jih postavi v besedni red ciljnega jezika in jih pregiba. Strojno prevajanje je najbolj učinkovito za besedila v t.i. nadzorovanem jeziku (jezik z vnaprej definiranimi besedišči in stavčnimi strukturami, ki zagotavlja kakovost in terminološko ustreznost strojnih prevodov; ang. *controlled language*) z ozkim semantičnim poljem, ki potrebuje le še naknadno preverjanje. Strojno prevajanje svojim uporabnikom daje možnost izbire in dopolnjevanja slovarjev.

7. Po prevajanju na ravni segmenta: Zaznava manjkajočih segmentov ter pregled oblike in slovnice. Program opozori na manjkajoče segmente, slovnične nepravilnosti in spremembe oblik.

8. Pregled nad prevajalskim postopkom in obračun: Ta posredni del prevajanja služi spremljanju napredka prevajalskih projektov. Vsebuje podatke o prevajalcu, rokih, spremembah besedil, prevajalskih prioritetah, datumih popravljanja itn. Takšna logistika spremljanja stopenj prevodov in obračunavanja je pomembna predvsem pri večjih in večjezičnih prevajalskih projektih.

3. RAČUNALNIŠKE TEHNOLOGIJE ZA PREVAJANJE IN TERMINOLOGIJO

Sanje o samodejnem prevajanju med ljudmi obstajajo že dolgo (predlog o mehanskem slovarju Descartesa in Leibniza seže v 17. stol.). Sistemi za samodejno prevajanje se tako razvijajo že od časov izuma elektronskega računalnika v štiridesetih letih. Dolga leta zatem se je tako prevajanje izvajalo neposredno prek dvojezičnih slovarjev in postopek je vključeval skopo analizo drugih jezikovnih prvin.

Že kmalu po izdelavi prvih računalnikov se je pojavilo zanimanje za računalniško obravnavo naravnega jezika. Najbolj odmeven je bil ameriški projekt avtomatskega prevajanja v 60-ih med angleščino in ruščino, posebej še, ko se je izkazalo, da so bile optimistične obljube o kvaliteti prevodov zelo pretirane.

V sedemdesetih letih so naredili prvo verzijo sistema SYSTRAN (AN-FR). V osemdesetih letih je napredek v računalniškem jezikoslovju prinesel bolj kompleksen pristop k prevajalskemu postopku. Ti sistemi so vključevali programe za prepoznavanje besednih oblik

(morfološka raven), strukture stavkov (skladenjska raven) in razpoznavanje večpomenskosti, homonimov ter leksikalnih razmerij (leksikalna raven).

Skozi zgodovino so bili razviti **trije tipi sistemov strojnega prevajanja**:

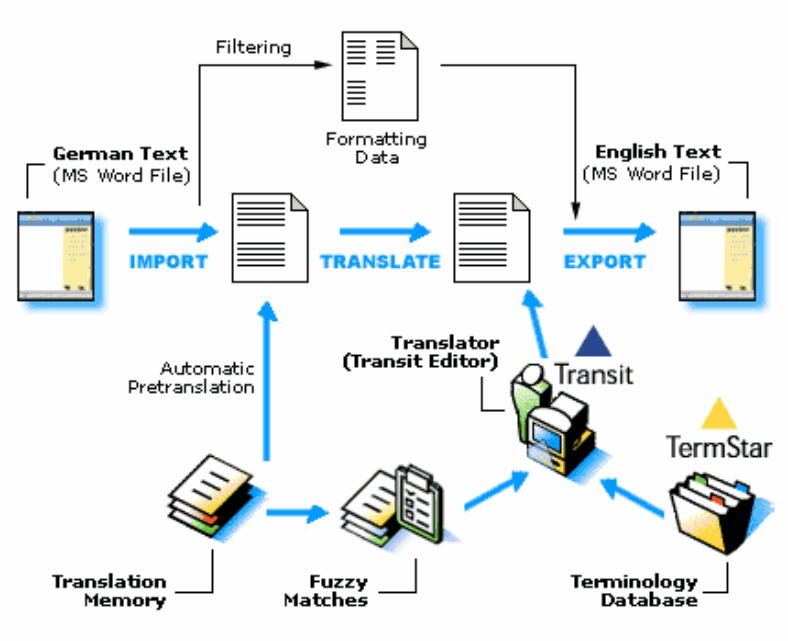
- Pri neposrednem pristopu, ki je najstarejši, gre za strojno prevajanje posameznega para jezikov v eno smer. Izhodiščno besedilo je analizirano zgolj za potrebe pretvarjanja v ciljni jezik.
- Vmesno stopnjo prevajalskega postopka predstavlja od jezikov neodvisni t.i. vmesni jezik (ang. *interlingua*). Pri tem gre za dva dela postopka: prevod iz izhodiščnega jezika v vmesni jezik, kateremu sledi prevod iz vmesnega jezika v ciljni jezik.
- Transferni pristop ima tri stopnje:
  - pretvorbo izhodiščnega besedila v abstraktno izhodiščno predstavo, kjer se razreši večpomenskost ne glede na jezik,
  - prenos predstave v abstraktno ciljno predstavo in
  - tvorbo besedila v ciljnem jeziku.

S pojavom osebnih računalnikov se je začel tudi razvoj prevajalske programske opreme zanje. Danes je razvoj usmerjen k statističnim sistemom, ki se prevajanja naučijo iz vzporednih korpusov (npr. EGYPT).

V devetdesetih letih se pojavijo komercialni prevajalniki in prvi spletni prevajalniki. V teh letih so postala priljubljena t.i. prevajalska namizja (ang. *workbench*), ki združujejo večino prevajalskih orodij (za štetje besed, pretvarjanje formatov, filtriranje, poravnavanje že prevedenega besedila). Tako so omogočala večjezično obdelovanje besedil, pošiljanje in sprejemanje dokumentov v elektronski obliki, pretvorbo grafičnih zapisov v besedila elektronske oblike, upravljanje terminologije s konkordančnim iskanjem, pomnilnike prevodov itd.

Prevajalska namizja so računalnik prevajalcem predstavila v povsem novi luči. Dobili so orodja, ki so jim omogočala širok razpon uporabe. Kot vedno pa je vrednost tehnologij odvisna od kakovosti dela. Pri strojnem prevajanju slovarji in terminologija zahtevajo trud, čas in denar, pomnilniki prevodov pa se zanašajo na zbirko uporabnih prevodov.

Štiri največja prevajalska namizja, ki jih poznamo danes, so **TRADOS**, **STAR** (Transit), **LinguaNet** (TranslationManager) in **LAN**T (Eurolang Optimizer).



Slika: Prevajalski postopek namizja Transit.

Do pred nekaj leti so ti sistemi tekli na velikih računalnikih (ang. *mainframe computers*) in so bili naprodaj za več milijonov dolarjev. S kasnejšim razvojem moči osebnih računalnikov in operacijskih sistemov *Unix* je postalo dostopnih veliko rešitev enake kakovosti in natančnosti in to po ceni, ki si jo prevajalci lahko privoščijo.

Poleg tega je na internetu dostopna celo brezplačna programska oprema, s katero se proti patentiranju bori projekt prostovoljcev, imenovan GNU (GNU's Not Unix). Posledica možnosti nalaganja programske opreme, za katero ni potrebno plačilo in ki se jo pod njihovimi pogoji (v nespremenjeni obliki in brezplačno) lahko ponuja naprej, sta širjenje in lokalizacija. Slovenski GNU skuša v duhu GNU ponuditi tiste tehnologije, ki so prosto dostopne in vezane na slovenski prostor: internetni črkovalni servis Primož Trubar, prazne besede slovenskega jezika (predlogi, vezniki, zaimki, pomožni glagoli itn.), navodila za prilagoditev nekaterih računalniških orodij slovenskemu jeziku (npr. kodni nabori, tezaver slovenskega jezika, oblikoskladenjski slovar) ter nenazadnje orodje za lokalizacijo programov v obliki pomnilnika prevodov SMART skupine za slovenjenje Linuxa (Košir, Peterlin in Erjavec 1998).

Področje prevajalskih tehnologij se deli na dve veji, ki se med seboj tudi povezujeta:

- **strojno prevajanje**
- **računalniško podprto prevajanje**

4. STROJNO PREVAJANJE

Prevajanje jezikov so se med prvimi intenzivno lotili v začetku devetdesetih let pri slovitem IBM-u, ki so razvili prvi sistem za t.i. ***statistično strojno prevajanje*** in postavili temelje za nadaljnja raziskovanja in izboljšave. Statistično strojno prevajanje sicer ni dalo takšnih rezultatov, da bi ga lahko neposredno uporabljali v vsakodnevni uporabi, pa vendar so bili rezultati skoraj nepričakovano dobri.

Strojno (podprto) prevajanje postaja eno izmed pomembnejših področji jezikovnih tehnologij. V času vsesplošne globalizacije je za ohranjanje določenega jezika obstoj takih sistemov posebej pomemben. To še zlasti velja za tako majhen jezik, kot je slovenščina. Velika večina prevajalnih sistemov je namenjena predvsem podpori prevajanju, kar pomeni, da so le v pomoč prevajalcu pri njegovem delu, še vedno pa je prevajalec tisti, ki besedilo dokončno prevede. Seveda ob izredno hitrem napredku strojne in programske opreme nekateri prevajalni sistemi postajajo tako dobri, še posebej na ozko specializiranih področjih, da lahko govorimo že kar o izključno strojnem prevajanju, saj poseg prevajalca pogosto sploh ni več potreben.

Pod izrazom »***strojno podprto prevajanje***« razumemo prevajanje, ki ga v prvi vrsti izvaja človek prevajalec ob podpori različnih računalniških orodij, med katerimi igrajo pomembno vlogo t.i. pomnilniki prevodov. Ti so jezikovno neodvisni, delujejo pa tako, da prevajalcu poskušajo pomagati na podlagi prevodov celih stavkov, ki jih je za kolikor toliko uspešno prevajanje treba prej vnesti v prevajalno bazo. In teh nikakor ni malo. Zato se za splošno prevajanje taka vrsta program zdi precej neuporabna, se pa toliko bolj uporabljajo na omejenih področjih, kjer se določeni stavki precej ponavljajo (npr. navodila za uporabo, zakonodaja). Pomnilniki prevodov imajo vgrajene posebne mehanizme približnega iskanja, brez katerih bi bili v večini primerov povsem neuporabni. Tako pa lahko prevajalcu tudi v primeru, da konkretnega stavka ni v prevajalni bazi, predlagajo stavek in njegov prevod, ki je po obliki oz. uporabljenih besedah najbližji. Prevajalec mora seveda prevod ustrezno spremeniti, kar pa je običajno še vedno precej hitreje, kot če bi moral stavek prevesti v celoti.

Strojno prevajanje poteka brez pomoči človeka prevajalca. Taki sistemi so jezikovno odvisni, njihov razvoj pa zahteva izredno veliko znanja in časa. Brez poprejšnjega učenja so sposobni prevesti poljubno besedilo, kar je velika prednost pri njihovi uporabi. V splošnem njihov prevod ni vedno najboljši. Zato so pregled in popravki s strani uporabnika še vedno zaželeni.

Prevajalne sisteme s področja strojnega prevajanja bi lahko razvrstili po različnih merilih. Če kot merilo vzamemo »globino« oz. način dela, na kakršnega vsebinsko prevajajo, bi jih lahko v grobem razdelili v štiri skupine:

1. **Prevajalni sistemi, ki le menjajo besedno obliko za besedno obliko**. Taki sistemi so za pregibne jezike, kakršen je tudi slovenščina, v splošnem neuporabni. V poštev bi lahko prišli le pri zelo omejenem področju prevajanja.
2. **Prevajalni sistemi, ki znajo vhodno besedili lematizirati** (poiskati osnovno obliko besede). Tak sistem bi brez kakršne koli dodatne logike npr. slovenščino že dal kolikor toliko dobre rezultate. Seveda pa v splošnem stavki ne bi bili posebno berljivi, ker prevajanje še vedno poteka na podlagi zamenjave besed za besedo, pa tudi sintaksa izvornega in ciljnega jezika se ne upošteva.
3. **Prevajalni sistemi, ki poleg morfologije upoštevajo tudi sintakso obeh jezikov**. Prevod takega sistema je v splošnem že kar lepo berljiv, ker pa sistem nima pomenske analize, prevod bistveno ne ustreza vedno najbolj.
4. **Prevajalske sisteme, ki prevajajo na podlagi morfološke, sintaktične in pomenske analize**. Ti dajejo pri prevajanju najboljše rezultate, a seveda niso nezmotljivi. Da bi pravilno »razumeli« stavek, ki ga prevajajo, zaradi boljše analize del informacij prenašajo iz prejšnjih stavkov, saj se pogosto zgodi, da se nek stavek nanaša na prejšnjega ali prejšnje stavke. Osnovna enota prevajanja so tako odstavki. Tudi rezultati prevajanja so običajno boljši pri daljših stavkih, saj tam prevajalni sistem lahko natančneje določi pomen posameznih besed kot pri krajših stavkih.

V splošnem se prevajanje izvaja v **treh osnovnih korakih**:
Najprej se z analizo poskuša čim bolj natančno ugotoviti, kakšen je pomen posamezne besede v izvornem stavku ter kako so te med seboj pomensko povezane. Včasih, še zlasti pri navidez preprostih stavkih se zdi, da analiza ni posebej zahtevna stvar. Vendar ni tako. Analizator mora najti in upoštevati vse mogoče pomenske in sintaktične možnosti, saj nikoli ne moremo natančno vedeti,katera od njih je prava: ali tista, ki je sicer najobičajnejša, ali katera druga.

Kot primer vzemimo stavek:

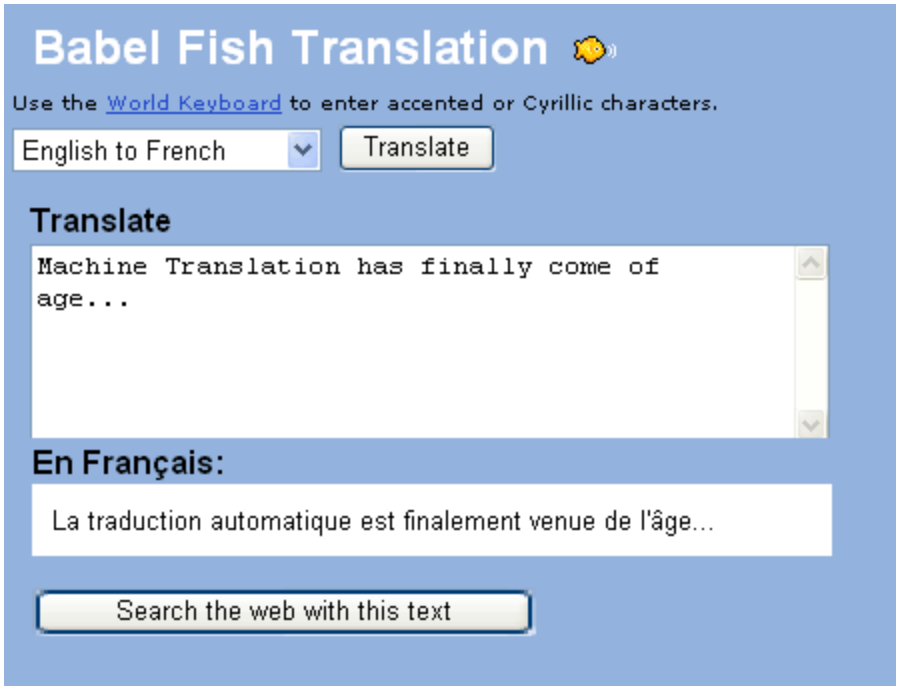
»*To je Miro Cerar*.«

Na prvi pogled ni nobenega dvoma. Pomen stavka si lahko razlagamo, kot da je to oseba moškega spola z imenom in priimkom Miro Cerar. Vendar pa si isti stavek lahko razlagamo še v dveh drugih pomenih, ki sicer nista običajna, sintaktično in pomensko pa sta povsem možna. Stavek lahko razumemo tudi tako, kot da to (nekaj) je žensko z imenom in priimkom Mira Cerar, ali pa, kot da oseba z imenom Miro Cerar je to (nekaj). Če bi imeli sintaktični analizator, potem so vse tri pomenske možnosti sintaktično pravilne in enakovredne. Šele z dodatno pomensko analizo, ki si pomaga z določenimi statističnimi informacijami, pridobljenimi s pomočjo analize posameznega jezika (korpusa), ter informacijami, pridobljenimi na podlagi prejšnjih stavkov, lahko te možnosti utežimo po verjetnosti ter s te določimo najverjetnejši prevod.

Ko tako izluščimo določen pomen posameznih besed ali fraz v izvornem stavku, moramo osnovne oblike teh besed ustrezno pretvoriti v besede in fraze ciljnega jezika. Pri tem si prevajalnik običajno pomaga z vgrajenim splošnim slovarjem. Ker ima lahko določena beseda ali fraza več možnih prevodov, je treba vse prenesti najprej do tretjega modula, ki se potem odloči za »pravi« prevod. Poleg osnovnega splošnega slovarja ima prevajalnik lahko še dodatne terminološke slovarje, osebne slovarje uporabnika ali vzorce prevodov celih stavkov, katerih naloga je čim bolje in s čim več možnimi besedami prevesti v ciljni jezik glede na področje, s katerega je prevajano besedilo. Vsako področje ima namreč svoje zakonitosti tvorjenja stavkov in specifične prevode določenih besed.

Ko je prevajalnik opremljen s podatki o pomenih posameznih besed in njihovih možnih prevodih, mora opraviti še tretjo fazo prevajanja, tj. **sintezo**. Ta ima nalogo, da oblikuje stavek po slovničnih pravilih ciljnega jezika, ki so lahko bistveno drugačna od pravil izvornega jezika. Poleg tega se mora odločiti za najboljši prevod ustreznega pojma. Pri tem lahko preprosto vrne tisti prevod, ki je v splošnem (na podlagi statistične analize) najpogostejši, lahko pa na podlagi drugih besed v stavku (na podlagi statistične analize kolokacij) določi prevod, ki je v primeru konkretnega stavka verjetno boljši. Vsekakor ima statistična analiza obeh jezikov, tako izvornega kot ciljnega, bistveno vlogo pri kakovostnem prevajanju. To je v bistvu znanje, ki se pri človeku (prevajalcu) odraža v obliki izkušenj, in na podlagi katerih se odloča tudi sam. Niso torej dovolj le veliki in dobri slovarji prevodov ter algoritmi za stavčno analizo in sintezo; prevajalnik mora vsebovati čim bolj popolno sliko o obeh jezikih, da se lahko kar najbolj približa dobremu prevodu.

Najnovejši prevajalni sistemi poskušajo kombinirati tako »čiste« prevajalnike kot pomnilnike prevodov. Tako dajejo uporabniku možnost, da po želji uporablja tiste mehanizme, ki so zanj najugodnejši glede hitrosti in kakovosti prevajanja. Hitrosti prevajanja, ki jih zmorejo sodobni prevajalni sistemi, so zelo odvisne od strojne in programske opreme, na kateri tečejo, navadno pa so sposobni prevesti nekaj deset strani (formata A4) besedila na minuto. To pa zlasti pri prevajanju večjih količin besedila seveda lahko prinese velikanske prihranke časa.
Težav, ki jih je treba rešiti pri tako zahtevnem opravilu, kot je prevajanje, je precej. Postopno izboljševanje različnih delov prevajalnega sistema, npr. slovarskega dela, ki že zaradi spreminjanja jezikov verjetno ne bo nikoli končano, pa bo prej ali slej prineslo rezultate, s katerimi bodo zadovoljni tudi najzahtevnejši uporabniki.



Slika. Primer strojnega prevajanja.

Kakšen je dober prevod, je kompleksno vprašanje, na katerega se ne da preprosto odgovoriti. Pri strojnih prevajalnikih je odločilen podatek, koliko sprememb in popravkov potrebuje prevod (čeprav zna prevajalnik tudi sam zaznati nekatere napake in sam ponuja načine, kako se izogniti napakam), da ustreza prevajalcu, bralcu ali naročniku, in koliko časa se pri tem porabi. Pred tem kriterijem pa se slovenski uporabnik strojnega prevajanja spopade še z večjo oviro. Širše uporabnih in prosto dostopnih strojnih prevajalnikov pri nas namreč še ni. Na

srečo so nekatera orodja za pridobivanje slovenskih jezikovnih virov in jezikovni viri drugih jezikov dostopni v tujini ali celo brezplačno na internetu.

Uporabnost sistemov strojnega prevajanja pa je odvisna tudi od drugih dejavnikov, med katerimi je treba posebej upoštevati izhodiščno besedilo samo. Da se izognemo nepotrebnim napakam, je priporočljivo izhodiščno besedilo najprej pripraviti (ali prenesti v nadzorovani jezik), pri čemer moramo paziti na enostavno stavčno strukturo in manjši obseg besedila, ki gre v prevod. Besedila ne smejo imeti napak pri črkovanju (te so lahko posledica nenatančnega skeniranja dokumenta), slovničnih napak, neslovničnih struktur in leksikalnih dvoumnosti.

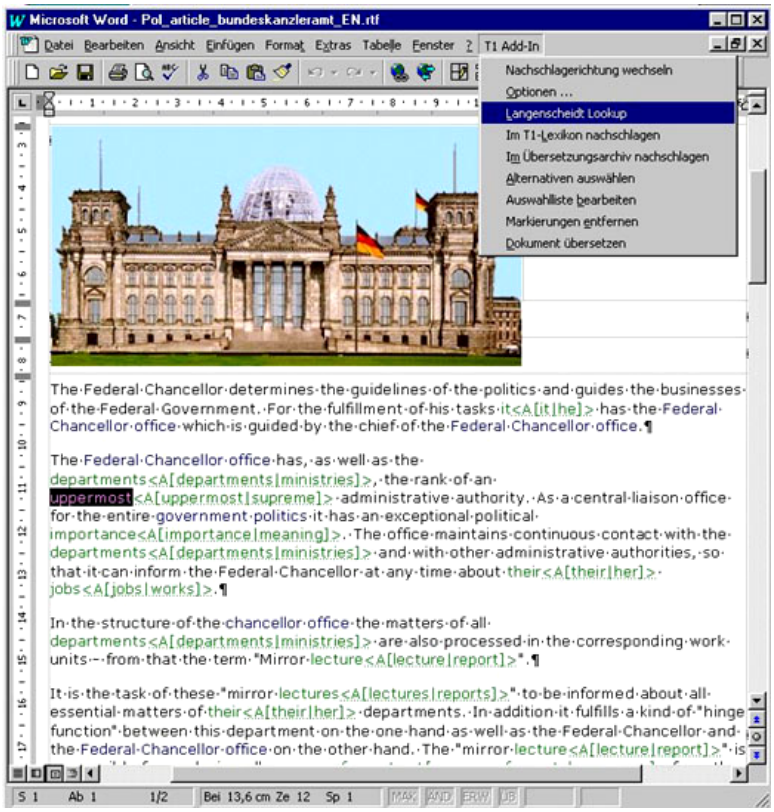
Prevajalniki so uporabni predvsem pri besedilih z določenega področja, za katera mora biti v računalnik vnešena tudi vsa potrebna terminologija (tudi žargonski izrazi), ki se v takšnih besedilih velikokrat ponavlja. Stavčna struktura takšnih besedil navadno ni zapletena oz. bi jo lahko lažje prilagodili.

Ob nepravi uporabi se lahko strojni prevajalnik res izkaže za zamudnega. Vseeno uporabniki počasi začenjajo sprejemati prednosti strojnega prevajanja pri enostavno strukturiranih besedilih, saj je uporabno tudi za preverjanje besedila.

Strojno prevajanje je dostopno tudi na **internetu**. Uporabniki lahko v določenem obdobju uporabljajo poskusne različice strojnih prevajalnikov ali pa kot stalni uporabniki izkoristijo strežniško ponudbo (*Systran*, *Logos*, *GLOBALink*). Dodaten znak velikega vpliva interneta je naraščanje programov za strojno prevajanje spletnih strani, elektronske pošte in pripetih dokumentov ter klepetalnic. Nekatera orodja lahko za uporabo strojnih prevajalnikov in pomnilnikov prevodov jezik izhodiščnega besedila priredijo nadzorovanemu jeziku.

Strojne prevajalnike uporabljajo tudi poslovni ljudje, in sicer za prevajanje glavnega pomena dokumentov, elektronske pošte, časopisnih člankov, poslovnih pisem; koristni so pri izbiri besedil, ki potrebujejo prevod strokovnjaka, za posredovanje informacij, ki spremljajo izhodiščno besedilo ter za izdelavo prevodov kot polizdelkov za nadaljnjo uporabo.

V zadnjem času se vedno pogosteje pojavlja težnja po združevanju strojnih prevajalnikov s pomnilniki prevodov. Translator's Workbench je danes na voljo z integriranim sistemom za strojno prevajanje podjetja Logos, ki vskoči pri vseh prevodnih enotah, ki nimajo ustreznic v pomnilniku prevodov. Obratno tudi vse več komercialnih sistemov za strojno prevajanje (npr. Langenscheidt) ponuja komponento za arhiviranje strojno prevedenih in popravljenih stavkov, ki (kot pomnilnik prevodov) služijo za referenco ob novih prevodih. Pri nas komercialni prevajalnik **PRESIS**, ki ima vgrajen tudi pomnilnik prevodov, prevaja zaenkrat iz slovenščine v angleščino).



Slika: Primer strojnega prevajalnika Langenscheidts T1 Professional 4.0.

Podjetja se zavedajo potenciala strojnega prevajanja, zato se v bližnji prihodnosti obetajo nove, izboljšane verzije in sveže ideje. V Singapurju, na primer, je bil že leta 1995 lokalno razvit sistem za prevajanje iz angleščine v kitajščino, malajščino, japonščino in korejščino, ki ga pregledujejo poklicni prevajalci. Sistem omogoča prevajanje ogromnih količin dokumentov za naročnike z vsega sveta, ponuja pa tudi lokalizacijo podjetjem, ki razvijajo programsko opremo za kitajsko govoreči del tržišča.

4.1 Zapletena slovenščina

Prevajanje iz enega jezika v drugi ni bil več problem samo za jezikoslovce, temveč so se reševanja zapletenega sporazumevanja lotili matematiki, statistiki, informatiki in še kdo. Osnova vsega so bile *obsežne zbirke besedil*, t.i. *korpusi*, ki so vsebovali poved v izvirnem jeziku in istočasno njen prevod. Računalniki so nato ugotavljali postavljanje posameznih besed v kombinaciji z drugimi in izračunali verjetnost tovrstnega ponavljanja. Iz tega sledi, da bi pri čim večji verjetnosti lahko »zadeli« pravi prevod. Ob tem še posebej ugotavljajo, kdaj se kakšne besede pojavljajo v kombinaciji z drugo, ko se prevod približa želenemu rezultatu, in kdaj se ne, ko je prevod morda nesmiseln.
Slovenski jezik je visoko pregiben, torej z velikim številom končnic besed z besednim repom, ki ni natančno določen. Ob tem imamo še slovensko značilnost, dvojino, ki je posebnost med večino slovanskih jezikov, in če se spomnimo na ure slovnice, se lahko večina samostalnikov tvori edninsko, dvojinsko ter množinsko, in to v šestih sklonih. Večina pridevnikov lahko tvori tri spole, vsa tri števila, šest sklonov in tri osnovne ravni stopnjevanja, ob tem pa ima slovenski jezik še mnoga izpuščanja. Sodobni in zmogljivi računalniki bi zagotovo bili kos tem osnovnim zahtevam, ki so kot naročeni za »umetno inteligenco«, vendar se osnovnim jezikovnim pravilom sintakse, gramatike, sinonimom besed oz. fraz priključi še zapletena semantika oz. pomen prevajalskih sporočil. Pot do uspešnega samodejnega oz. stojnega prevajanja (machine translation) iz enega v drug jezik, bo zagotovo še dolgotrajna.

4.1 Prevajalni sistem PRESIS

Prevajalni sistem Presis je namenjen strojnemu prevajanju besedil. Za zdaj sta izdelana prevajalnika med slovenščino in angleščino in obratno. Zaradi vgrajenega pomnilnika prevodov je primeren za različne skupine uporabnikov. Presis je hitra in učinkovita pomoč pri prevajanju, saj namesto vas opravi celo vrsto rutinskih opravil, z dodatnimi nasveti pa vam pomaga pri oblikovanju končnega prevoda. Zaradi upoštevanja slovničnih pravil slovenskega in drugih jezikov je Presis učinkovitejši od klasičnih prevajalskih orodij, še posebej v kombinaciji z elekronskimi slovarji, ki jih je razvilo podjetje Amebis.

Slika: Prevajalni sistem Presis.

Primer uporabe prevajalnika Presis:

Če želimo preizkusiti kakovost prevajalnega sistema Presis, v polje vpišemo svoje besedilo, ki ga želite prevesti.

Sem vpišite besedilo, ki ga želite prevesti.

Besedilo za prevajanje

156

Prostih znakov

angleščina >> slovenščina

slovenščina >> angleščina

Prevajalni sistem Presis ponavadi ne prevaja povsem pravilno. Prevod je namenjen približnemu razumevanju originalnega besedila ali osnovi za nadaljnje prevajanje. Vendar je kakovost prevoda odvisna tudi od dejavnikov, na katere lahko vplivate. Za boljše prevajanje se moramo držati navodil.

- JavaScript in piškotki (Cookies) sta nujna za pravilno delovanje teh strani.
- Stavke in lastna imena moramo pisati z veliko začetnico (imena jezikov, dni v tednu in mesecev v slovenščini niso lastna imena).
- Besedilo za prevajanje morajo biti slovnično pravilna, saj to pomaga pri boljši analizi.
- Predstavitvena verzija prevajalnega sistema Presis je omejena na prevajanje besedil z dolžino največ 200 znakov. V predstavitveni verziji ni mogoče dodajanje lastnih prevodov.

Po začetni fazi, ko je bilo prevajanje večinoma na nivoju besed z nekaj osnovne stavčne analize, je *Presis* zdaj prešel na uporabo glagolskih predlog pri stavčni analizi in generacij in prevodov.

Predloga opiše, kako se določen glagol uporablja, pove, s katerimi predmeti se veže, katere omejitve so pri tem, katera prislovna določila so lahko ob glagolu, katere predložne zveze so tipično povezane z glagolom in podobno. Predloge so potem povezane v pomen, pri čemer so pripisane tudi potrebne stilne oznake, ki jih prevajalnik uporablja za boljše izbiranje prevodov. Da bi se izognili prevelikemu ponavljanju pri predlogah, je del informacij, ki so značilne za veliko število glagolov v posameznem jeziku, rešen programsko na nivoju prevajalnika (analizatorja oz. generatorja) in so v teh primerih v predlogah poudarjene le izjeme. Tak primer je npr. vezava s prislovnimi določili, kjer so tipično dovoljena dodatno k predlogu poljubna prislovna določila, razen prislovnega določila kraja, po katerem se vprašamo kam. Analizator in generator tudi skrbita za pravilen vrstni red prislovnih določil v posameznem jeziku.

S pomočjo predlog lahko razrešimo pri prevajanju dvoumnost, ki se slovnično razlikujejo, recimo glede prehodnosti/neprehodnosti, povratnosti pri slovenskih glagolih, frazne glagole v angleščini,… Prav tako je možno stalne fraze v celoti vpisati v predloge, s čimer lahko dosežemo pravilne prevode fraz. Ta zapis je veliko bolj prilagodljiv, kot bi bil zapis z enostavnim zaporedjem besed, saj so elementi predloge lahko v poljubnem vrstnem redu (znotraj pravil posameznega jezika).

Tako lahko Presis zdaj brez težav prevede tudi stavek:

»Bobu je včeraj Janez rekel bob.« v *»Janez called a spade a spade yesterday.«*

Pri tem sicer najde tudi druge možne prevode (brez uporabe fraz), vendar da na prvo mesto rešitev, kjer predloga pokrije čim večji del stavka (postopek za razvrščanje najdenih analiz uporablja sicer več kriterijev za razvrstitev, vendar je ta med pomembnejšimi in ponavadi prevlada).

Predloge določijo tudi v stavčne člene, tako da je zaradi tega mogoče povezovati pomene glagolov s pomeni samostalnikov glede na vlogo, v kateri se pojavljajo. Tako je mogoče določiti, da sta glagol (oz. glagolski pomen) »skuhati« in samostalnik »kosilo« povezana tako, da je kosilo tipično predmet v tožilniku (oz. rodilniku v zanikanih stavkih, za kar poskrbi sam analizator in tega v vzorcih ni treba posebej pisati) pri tem glagolu, ne pa npr. osebek ali predmet v dajalniku.

Pri predlogah je možno napisati tudi tipične predložne zveze, s katerimi se povezujejo. S tem lahko po eni strani izboljšamo prepoznavanje predlog, po drugi strani pa tudi dosežemo, da se predlog ob določenem glagolu prevaja drugače, kot bi se sicer običajno.
»*V našega Janeza se je zaljubila lepa Micka*.« tako Presis prevede v »*Beautiful Micka fell in love with our Janez*.«

Presis zna predloge tudi prilagoditi različnim časom in trpniku. Morebitne izjeme (predloge, ki ne morejo uporabljati v trpniku ali se lahko uporabljajo le v trpniku, omejitve pri časih) so označene z uporabo lestvic pri predlogah. Če stavka, ki ji bil v originalu v trpniku, v ciljnem jeziku ni mogoče zapisati na ta način (to se običajno zgodi pri prevajanju iz angleščine v slovenščino, kadar je znan tudi osebek, ki je bil v angleščini naveden z »by«), ga generator pretvori v tvorni način. Primer za to je prevod iz:
»*John was made happy by Mary*.« v »*Mary je osrečila Johna*«.
Posebnost slovenščine so tudi glagoli, ki se dobijo v neosebni rabi drug pomen. Tak primer je:
»*Gre za ta problem*.«, kar Presis s pomočjo predloge prevede v
»*It is this problem*.« oz. kot druga možnost
»*It is about this problem*.«
Najde pa seveda tudi kopico drugih možnosti, recimo
»*It goes behind this problem*.«

Predloge so lahko take, da kot del zahtevajo nov glagol. Tipičen primer so predloge za modalne glagole (oz. v slovenščini tudi modalne prislove, recimo »lahko« in »rad«). V takem primeru zna Presis sestaviti predlogi. Primer za tak prevod je:
»*Why could John fall in love with Mary*?«, kar prevede v
»*Zakaj bi se John lahko zaljubil v Mary*?«

Ocenjujejo, da bo za zadovoljivo delovanje prevajalnika potrebnih med 10.000 in 20.000 vnesenih predlog na jezik, jezik pa bo večinoma pokrit pri okoli 100.000 predlogah. Težava pri ugibanju predlog so v slovenščini povratni glagoli in glagoli v neosebni rabi, v angleščini pa frazi glagoli, zato imajo ti pri vnosu predlog prednost (skupaj z najbolj pogostimi glagoli).

5. RAČUNALNIŠKO PODPRTO PREVAJANJE

Računalniško podprto prevajanje (ang. CAT – *Computer-Aided Translation*) predstavlja drugo vejo prevajalskih računalniških tehnologij, ki se je razvila s prevlado osebnih računalnikov.

Uporaba teh orodij olajšuje in pospešuje, optimizira in poceni prevajalski postopek in ga ne simulira kot strojno prevajanje. Ti programi nam služijo za podporo referenčnega dela, tj. iskanja po slovarjih, vzorčnih besedilih, terminoloških bazah. V to skupino orodij spadajo elektronski slovarji, črkovalniki, programi za preverjanje slovnice, slovarji sopomenk, terminološke baze, pomnilniki prevodov in drugi računalniški podatkovni viri.

5.1 Pomnilniki prevodov

Po definiciji skupine strokovnjakov za standarde jezikovnega inženiringa EAGLES (*Expert Advisory Group on Language Engineering Standards*) je pomnilnik prevodov »večjezični besedilni arhiv, ki vsebuje (segmentirana, poravnana, razčlenjena in klasificirana) večjezična besedila in dovoljuje shranjevanje besedil in iskanje po njih glede na različne pogojev. Pomnilnik prevodov je podatkovna zbirka prevodnih enot, navadno povedi ali krajših delov besedila, ki so v izvirniku in prevodu shranjeni v pomnilnik in so ob morebitni ponovitvi enakega ali zelo podobnega dela besedila na razpolago za ponovno uporabo.

Pomnilnik prevodov je lahko integriran v urejevalnik besedil, lahko pa ima lastno delovno namizje, v katerega uvozimo dokument, ki ga želimo prevesti. Navadno obsega še orodje za izdelavo in upravljanje terminoloških enot, komponento za vzporejanje, s katero pomnilnike ustvarjamo iz že prevedenih besedil, preverjanje črkovanja, strojno prevajanje, lahko pa ima tudi statistični program, s katerim lahko ugotovimo t.i. faktor ponavljanja v besedilu. Ta nam pove, kako pogosto pride do ponovitev, kar nam je v pomoč pri izbiri primernega prevajalskega postopka in orodja.

To orodje nima vgrajenih modulov za oblikoskladenjsko analizo prevodnih enot, niti lastnih leksikonov, saj deluje na jezikovno neodvisnem principu, kar pomeni, da ne zaznava podobnosti pomenov. Sposobno je prepoznavati podobnost na ravni besed ali besednih nizov, zato je uporabno za vse jezikovne pare oz. za vse jezike, za katere je zagotovljena znakovna

podpora. Program med prevajanjem v ozadju išče enake (popolni zadetek, ang. *exact match*) ali podobne enote (megleni zadetek, ang. *fuzzy match*), ki jih prevajalcu samodejno ponudi. Podobnost je odvisna predvsem od števila besed, ki se ujemajo v obeh prevodnih enotah, in besednega reda. Prag ujemanja lahko določi prevajalec sam.
Največkrat uporabljani programi s pomnilnikom prevodov so **TRADOS Translator's Workbench**, **ATRIL DéjaVu** in **STAR Transit**.

5.2 Terminološki programi

Spreminjajoče se terminologije, zaradi nenehnega razvoja strokovnih in drugih področij, pogosto ne more spremljati ne slovaropisje ne prevajalec. Prevajanje besedil s področij kot so proizvodnja, energija, pravo, medicina idr. je lahko zato zelo naporno, saj je iskanje izrazov in njihovih prevodov lahko dolgotrajno in neuspešno. Veliko izrazov je moč najti na internetu in v drugih javnih medijih, terminologija pa je lahko v lasti izdelovalca terminološke baze oz. naročnika prevoda in tako zaščitena z avtorskimi pravicami. Prevajalec ali skupina prevajalcev, ki večinoma prevajajo besedila določenega področja, si zato sami ustvarjajo terminološko bazo, ki jim v naslednjih prevodih zagotavlja tudi enotnost pri izbiri izrazov. Delajo pa lahko tudi na različnih jezikih, saj lahko terminološki program za posamezen izraz shranjuje večjezične prevodne ustreznice.
Terminološki programi so orodja za izdelavo in vzdrževanje terminologije. Imajo vlogo skladišča, kamor se zbirajo in shranjujejo izhodiščni in ciljni izrazi za kasnejšo uporabo v prevodu. Hranijo lahko neomenjeno število terminoloških vnosov. Tehnike shranjevanja in prikazovanja izrazov pa so različne od programa do programa. Ta (lahko) vsebuje orodja, ki:

- strukturirajo, posodabljajo in povezujejo vnose,
- omogočajo preproste funkcije iskanja,
- omogočajo konceptualni prikaz popolnih in meglenih zadetkov,
- podpirajo shranjevanje grafičnih prikazov,
- omogočajo samodejno vnašanje izrazov v urejevalnik besedil
- z jezikovno analizo izhodiščnega in ciljnega besedila prepoznajo in izločijo izraze za uvoz v terminološki program,
- vključujejo tudi slovarsko upravljanje terminologije,
- podatkovno bazo izvozijo in uvozijo v druge aplikacije.

V nekaterih pogledih so zelo podobni pomnilnikom prevodov:

- Podpirajo vse jezike, za katere je zagotovljena znakovna podpora, saj je iskanje tudi tu pogojeno s podobnostjo besed.
- Omogočajo globalno iskanje (iskanje tudi po delih izraza), megleno iskanje (prikaže se kazalo besednih zvez, ki poleg korena iskane besede vsebuje tudi tvorjenke, oblikoslovne različice besed ipd.) in filtriranje (prikaz vnosov po kriterijih, kot jih določi uporabnik).
- Omogočajo doslednost in enotnost.
- Terminološki vnosi so opremljeni s podatki o vnašanju (vnašatelj, datum vnosa, datum spremembe, področje, kje v besedilu se izraz nahaja itn.) in o izrazu samem (o rabi, obliki, lastnostih, definiciji idr.).
- Terminološka baza je ob nakupu prazna in neuporabna, dokler vanjo ne vnesemo terminoloških vnosov. Čas, ki ga potrebujemo za vnašanje izrazov, lahko skrajšamo s predpripravo enojezične baze, ki ji samo dodamo prevodne ustreznice.
- Prevajalec lahko uporablja terminološko bazo kot dopolnilo pomnilniku prevodov ali drugim jezikovnim virom. Podjetjem, ki se resno ukvarjajo s prevajanjem, pomeni takšen terminološki program dragocen jezikovni vir.

Največkrat uporabljani terminološki programi so **TRADOS MultiTerm**, **ATRIL Terminology Management** in **STAR TermStar**.

6. KORPUSI

Korpus je zbirka besedil, ki so izbrana tako, da karakterizirajo stanje ali raznovrstnost nekega jezika. Uporaben je kot osnova, na kateri gradimo opise jezika, ali pa kot sredstvo za preverjanje hipotez o jeziku. Korpusi so dandanes že standardno shranjeni na računalnikih, saj ti po eni strani omogočajo kompaktno in poceni hranjenje ter razširjanje velike količine besedil, po drugi strani pa besedila lahko z njimi bolj učinkovito izkoriščamo.

Obdelava korpusov, takrat večinoma še v papirnati obliki, je bila v veliki meri prisotna že v petdesetih in šestdesetih letih. Zaradi Chomskyjeve podpore preučevanja »notranjega jezika« oz. človeške sposobnosti produkcije jezika in njegovih drugih vplivnih teorij je zanimanje za korpuse za nekaj let zbledelo in se spet prebudilo v osemdesetih, predvsem zaradi hitrega

razvoja najrazličnejših tehnologij, empirične narave raziskovanja in povečanja količine besedil ter kakovosti korpusov.

Osnovni namen korpusov je omogočanje temeljitega vpogleda v jezik na najrazličnejših ravneh in področjih. Tako jih lahko s pridom uporabljamo v jezikoslovju, v humanističnih in družboslovnih vedah in celo v informatiki in matematiki. Večinoma so uporabni v leksikologiji in predvsem leksikografiji, zdaj pa se z različnimi tipi, ki so dostopni širšemu krogu ljudi, vse bolj širijo na vsa jezikoslovna področja. Koristno jih torej lahko uporabljamo v slovaropisju, pri jezikovnih študijah, razvoju jezikovnih tehnologij, pa tudi za dinamična in z gospodarstvom neposredno povezana področja jezika, kot je terminologija. Z njihovo pomočjo lahko sestavljamo terminološke slovarje, odkrivamo že uporabljene izraze, prevode in razlage, s čimer je izdelovanje slovarjev ažurnejše in cenejše.

Korpusi so zgrajeni po različnih kriterijih. Pomembni so zunajjezikovni dejavniki, kot so medij, slog, žanr, datum publikacije itn. Vzporedno z gradnjo velikih računalniških korpusov se gradijo tudi vse boljša računalniška orodja za njihovo označevanje, analizo, upravljanje in iskanje po njih.

Programi za delo s korpusi so pregledovalniki oziroma konkordančniki, ki so sposobni poiskati željene dele korpusa in informacije ustrezno predstaviti. Najbolj znana orodja na internetnem tržišču so *Wordsmith*, *MonoConc* (za enojezične korpuse) in *ParaConc* (za vzporedne korpuse).

Uporabnost nekega korpusa je odvisna od njegove velikosti pa tudi urejenosti, tj. kako podrobno je dokumentiran in označen, ter standardiziranosti njegovega zapisa. Koristnost korpusa je nedvoumna, a njegova izdelava, razširjanje in uporaba so razmeroma zahtevni. Izdelava korpusa pa je smiselna le, če se ta tudi uporablja. Lahko je zelo draga, dodatne težave pa lahko povzroča še pravno vprašanje, kdo je njegov lastnik (avtorji, založbe, prevajalci itd.)

Delitev korpusov po nekaterih tipih glede na uporabnost pri prevajanju in v prevodoslovju:

- **Vzporedni korpusi** (izvirna besedila in njihovi prevodi) so nepogrešljivi pri prevajanju, za izdelavo prevajalskih pripomočkov, pri izboljšavah za strojno prevajanje in za izdelavo terminoloških baz, kot orodje pri programih za poučevanje

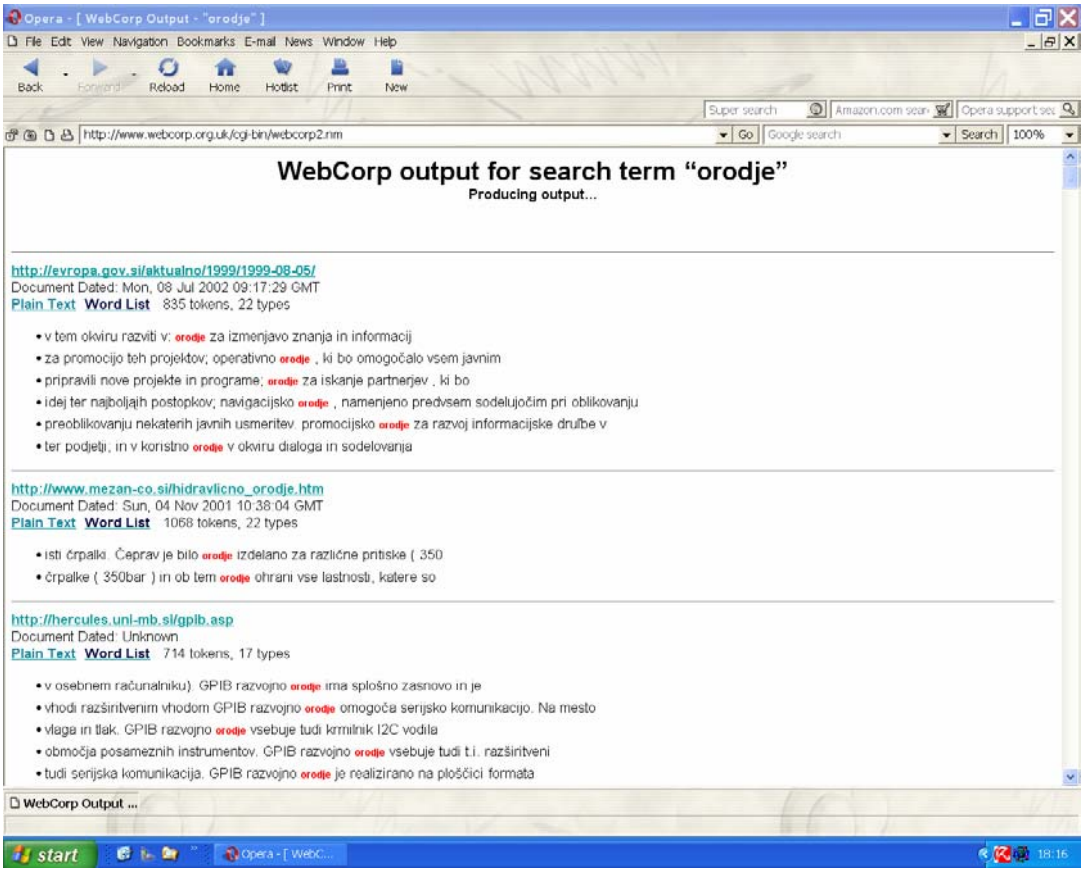
prevajanja, pri učenju jezikov s pomočjo jezikovnih tehnologij in pri terminoloških študijah, kjer so terminološki izrazi izluščeni iz korpusov. Največji uporabniki vzporednih korpusov so Združeni narodi, Nato, Evropska unija in države z dvema uradnim jezikoma (npr. Kanada). V vzporednih korpusih lahko iščemo prevodne ustreznice s pomočjo vzporednih konkordanc. Primerjamo lahko pogostost posameznih prevodnih ustreznic in njihova sobesedila, kar olajša izbiro primernega prevoda.

- Z **večjezičnimi korpusi** (nizi dveh ali več enojezičnih korpusov v različnih jezikih, izdelanih na podlagi podobnih kriterijev) dostopamo do naravnih vzorcev v jeziku, saj nam nudijo vpogled v jezikovne strukture v njihovem domačem okolju in ne v prevedenem besedilu. S prepoznavanjem strokovnih izrazov oz. njihovih prevodov, besed in fraz se približujejo pomnilnikom prevodov in terminološkim bazam. Pomembno vlogo imajo pri materialih za pisanje, poučevanju prevajalcev in pri izboljšavi programov za strojno prevajanje.
- **Primerljivi korpusi** (križanci med večjezičnimi in vzporednimi korpusi) so sestavljeni iz dveh posameznih ločenih zbirk besedil v istem jeziku: iz besedil v izvirnem jeziku in iz zbirke prevodov v ta jezik iz enega ali več drugih jezikov (npr. časopisni članki iz evropskih časopisov v nekem obdobju). S tem je možno prepoznavanje vzorcev, ki so specifični za prevedena besedila ne glede na izhodiščni oz. ciljni jezik, kar sproži nove hipoteze o postopku prevajanja, ugotavljanje prevodnih norm v specifičnih kontekstih ter odkrivanje metod in rešitev za poklicne prevajalce.

  Primer vzporednih in primerljivih korpusov je korpus MULTEXT-East, ki zajema šest srednje- in vzhodnoevropskih jezikov (med njimi tudi slovenskega) in je nadaljevanje projekta MULTEXT šestih jezikov Evropske unije.
- Tako je korpus izvirnih besedil pravzaprav **enojezični korpus**, ki je prav tako uporaben kot pomoč študentu prevajanja pri razumevanju nematernega jezika in razvijanju sposobnosti izražanja v maternem.
- Za prevajalce so lahko uporabni tudi **referenčni korpusi** (osnovna zvrst korpusov). Ti služijo kot jezikovni standardi, predstavljali naj bi idealizirano podobo jezika. Kot nasprotje referenčnim korpusom stojijo specializirani (služijo nekemu namenu) in oportunistični (cenena različica referenčnih korpusov; zbrani so glede na dane možnosti) ali spremljevalni korpusi (dinamični korpusi, v katerih je vidno spreminjanje jezika).

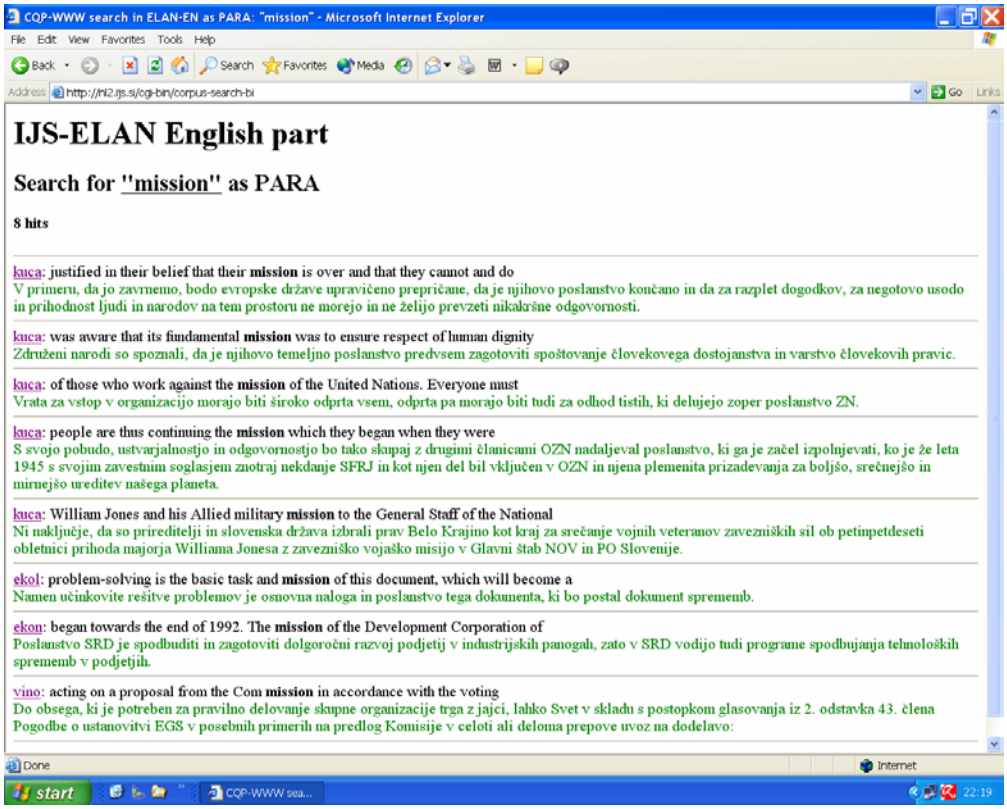
Največ raznovrstnih korpusov je za angleški jezik. Referenčni korpus angleškega jezika *British National Corpus* in spremljevalni korpus *Bank of English* sta dva največjih. Po njiju lahko iščemo od posameznih besed do daljših besednih zvez, do neke meje lahko celo določimo besedilno vrsto. Računalniški korpusi besedil so dandanes zelo priljubljeni tudi v Evropski uniji (*PEDANT*, *Intersect*).

V vlogi največjega svetovnega korpusa nastopa internet (ogromno podatkov, najhitrejše posodabljanje), vendar z določenimi pomanjkljivostmi (različno zastopane besedilne vrste – največ je besedil o računalništvu, nereprezentativnost). Za iskanje besed in besednih zvez ter večjezikovnih izrazov lahko uporabimo spletne iskalnike kot so *Google* (išče tudi po slovenskih izrazih), *Altavista*, *Yahoo* itd.. Pri tem moramo upoštevati, da bo iskalnik našel uporabne rezultate predvsem za dokaj redke besede. Med iskalniki so za prevajalca učinkovitejši tisti, ki mu poleg z naslovi spletnih strani postrežejo tudi s kratkim povzetkom vsebine. Eden najnovejših korpusnih iskalnikov na spletu je *WebCorp*, ki se pri iskanju besed poveže z spletnim iskalnikom (izbere ga uporabnik sam) in nam postreže s pravo konkordanco, obkroženo s sobesedilom.



Slika: WebCorp; Internet kot korpus

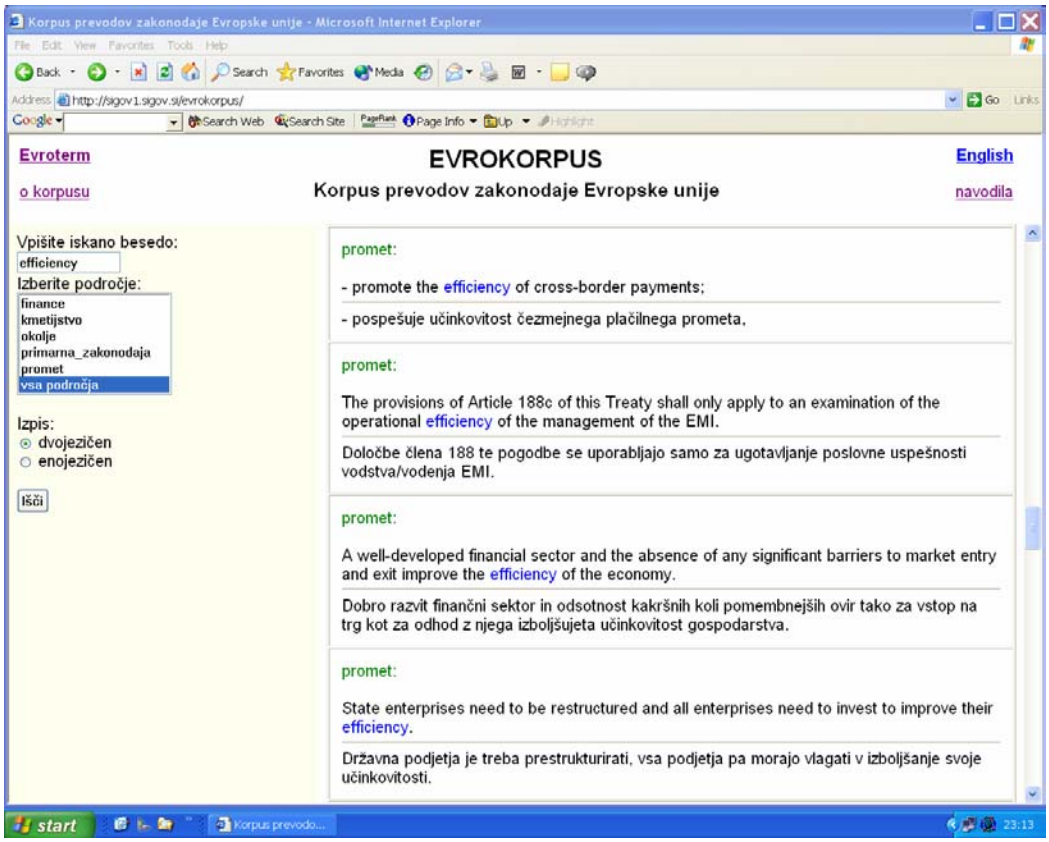
Slovenščina je jezik z malim številom govorcev. Vseeno pa je zanj, sicer brez vladnega financiranja, s sodelovanjem založb, računalniških hiš in akademskih institucij pred kratkim prišlo do večjega premika z enojezičnim referenčnim korpusom FIDA in dvojezičnim vzporednim korpusom ELAN (skoraj polovico besedil je prispeval Sektor za prevajanje SVEZ). Za to področje sedaj skrbi Slovensko društvo za jezikovne tehnologije.



Slika: Dvojezični vzporedni korpus ELAN.

V zadnjem času se izgradnja jezikovnih virov in tehnologij pospešuje. Korpusi so postali resnično uporabni šele v zadnjih letih, ko jih je vse več javno dostopnih prek interneta. Lahko jih uporabljamo tudi za lastne potrebe, bodisi s pomočjo orodij, ki jih nudijo programi s pomnilnikom prevodov, bodisi z že naštetimi orodji.

Evrokorpus je zbirka jezikovnih dvojic izvirnih in prevedenih besedil. Vsebino pomnilnikov prevodov so pretvorili v tekstno obliko in jo grupirali po področjih. Evrokorpus vsebuje več kot 1,5 milijona besed, ki pa še niso povsem prečiščene, saj je korpus šele v testni obliki.



Slika: Evrokorpus.

6.1 Korpus slovenskega jezika FIDA

Korpus slovenskega jezika FIDA je referenčni korpus za slovenski jezik; je rezultat projekta dveh pedagoško-raziskovalnih in dveh komercialnih partnerjev: Filozofske fakultete Univerze v Ljubljani, Instituta Jožef Štefan, založbe DZS, d.d. in podjetja Amebis, d.o.o. Projekt gradnje korpusa FIDA se je začel spomladi leta 1997, končan je bil ob koncu leta 2000.

Korpus slovenskega jezika FIDA je:

- **Referenčni korpus:** referenčni korpus je obsežna elektronska besedilna zbirka, ki zajema vzorčni delež besedil nekega jezika. Njegov osnovni namen je, da omogoča temeljit vpogled v jezik na najrazličnejših ravneh in področjih, in je tako pomemben vir za uporabno in teoretično jezikoslovje, npr. slovaropisje v vseh oblikah (eno- in večjezikovni slovarji, terminološki slovarji in drugi jezikovni priročniki), poučevanje jezika (učbeniki in učni pripomočki), jezikovne tehnologije (črkovalniki, slovnični pregledovalniki,

govorni vmesniki) ter tudi druge družboslovne in humanistične vede, npr. literarno vedo, psihologijo in sociologijo.

- **Enojezikovni korpus:** vključuje sodobna slovenska besedila; tujejezični elementi se v korpusu lahko pojavijo le kot sestavni del slovenskega besedila, izključena pa so vsa tujejezična besedila, npr. italijanska iz dvojezikovnih medijev na Obali.

- **Sinhroni korpus:** korpus sodobne slovenščine druge polovice 20. stoletja, vendar s poudarkom na zajemanju besedil, nastalih v 90-ih letih.

- **(Izhodiščno) pisni korpus:** zajema pisna besedila in prvotno pisna besedila, namenjena govorjenju.

V korpusu slovenskega jezika FIDA so zbrana sodobna slovenska besedila v skupnem obsegu nekaj nad 100 milijonov besed; v njem je zajeta široka paleta variant slovenskega jezika, kot ga prinašajo predvsem slovenski tiskani mediji, nekaj je tudi internetskih besedil in transkripcije govora

7. STANDARDI ZAPISA JEZIKOVNIH PODATKOV

Standardiziran računalniški zapis jezikovnih podatkov poveča uporabnost jezikovnih podatkov, saj poleg izmenljivosti spodbudi tudi njihovo večnamenost ter podaljša njihovo trajnost. Do prvih pobud za standardizacijo je privedel premik pri zbiranju in obravnavi jezikovnih podatkov, ki se je zgodil predvsem na področju računalniškega jezikoslovja oz. jezikovnih tehnologij in v podjetjih z velikimi količinami besedil.
Naloga standardizacije je predpisovanje javno dostopnih in trajnih načinov zapisa. Zapisi morajo biti podrobno definirani in shranjeni v enotnem formatu, če se hočemo izogniti težavam, ki se lahko pojavijo že pri zapisih črk.
Industrijski standardi so načini zapisa, ki so sicer v lasti nekega podjetja, a se uporabljajo tudi s programi drugih proizvajalcev, vsaj tako, da omogočajo uvoz in izvoz podatkov v tem formatu. Ti podatki so vezani na orodje, s katerim so nastali, obenem pa hitro zastarajo.
Mednarodni standardi pa so javni, večinoma prosto dostopni. Spreminja se jih po samo točno določenem postopku. Vendar pa je potrebna izbira in implementacija standarda za naše potrebe navadno zapletena in draga. Poleg tega je zaradi hitro razvijajoče tehnologije težko

vedeti, kateri se bodo obdržali. Pri teh standardih je treba upoštevati tudi mednarodna priporočila, ki jih je treba aplicirati in prilagoditi za slovenski jezik in za konkretne vire.
Ti standardi so:

- **SGML** (*Standard Generalized Markup Language*) podaja metajezik, ki služi za opis (pretežno) besedilnih dokumentov. Kot standard je bil sprejet že leta 1986 in ima bogato zgodovino uporabe. Določa jezik za predstavitev dokumentov, nad katerimi bodo delovali programi za obdelavo besedil. Eden od osnovnih ciljev SGML je, da so v njem zapisani podatki prenosljivi z ene strojne in programske opreme na drugo brez izgube informacij. Vedno več podjetij, ki imajo opravka z velikimi količinami besedil, prehaja na zapis SGML in vedno več podjetij se ukvarja izključno z izdelovanjem programske opreme ali s pomočjo končnim uporabnikom, da preidejo na ta standard. SGML služi kot osnova množici izvedenih standardov in mednarodnih priporočil.

- Medtem ko je **HTML** (*Hypertext Markup Language*) kot trenutni standard zapisa spletnih strani samo določen tip dokumentov SGML, je XML (eXtensible Markup Language), sicer še vedno poenostavljen SGML, metajezik za ustvarjanje informacijsko bogatih dokumentov in način za izmenjavo sporočil. Ta jezik je izmenljiv, odporen na tehnološke spremembe in omogoča uporabo dokumentov v različne namene. Zaradi zapletenosti standarda SGML in zaradi vse večjega pomena mrežne izmenjave podatkov je XML postal osnova za množico izvedenih standardov in pobud za zapis različnih zvrsti jezikovnih, pa tudi drugih strukturiranih podatkov.

- **TEI** (*Text Encoding Initiative*) so priporočila, ki jih upošteva večina projektov, ki zbira jezikovne vire. To so priporočila za pripravo in izmenjavo besedil za raziskovalne in založniške namene. Določajo konkretne oznake SGML in strukturo teh oznak. Priporočila TEI so zaenkrat najbolj natančno izdelani tip dokumentov SGML, ki pokriva raznovrstna gradiva (leposlovje, slovarji, zbirke besedil idr.) ter različne načine dodatnega označevanja teh gradiv (jezikovno, uredniško itn.). Na TEI se dandanes že samoumevno sklicujejo projekti, ki ustvarjajo jezikovne vire, predvsem korpuse. S TEI se povezuje tudi večina standardov, izvedenih iz SGML. Tudi TEI je svoja priporočila v zadnjem času preoblikovala v skladu z jezikom XML.

- **MARTIF** (*Machine Readable Terminology Interchange Format*) je zvrst dokumentov SGML, ki naj bi standardizirala računalniški zapis terminoloških baz. Primer takšne baze je Eurodicautom. Pretvorba večjega števila terminoloških baz v ta format omogoča iskanje večjega števila uporabnikov in ponovno uporabo vira za druge programe jezikovnih tehnologij.

- **TMX** (*Translation Memory eXchange*) je zvrst dokumentov SGML/XML, ki naj bi standardizirala računalniški zapis pomnilnikov prevodov. Za zapis pomnilnikov prevodov trenutno prevladujejo industrijski standardi (izdelki Trados in še širše, Word ter Microsoft), postopek standardizacije pa je po začetnem obotavljanju v zadnjih dveh letih zajel skoraj vse ponudnike prevajalskih orodij. Pretvorba pomnilnikov prevodov v ta format omogoča izrabo in izmenjavo večjezičnega vira besedil za izdelavo boljših in bolj ažurnih slovarjev z luščenjem izrazov ter za strojno prevajanje.

8. ZAKLJUČEK

Jezikovne tehnologije so področje računalniškega jezikoslovja, ki se ukvarja z jezikom kot merljivo zbirko udejanjenih primerov rabe z drugačnimi raziskovalnimi metodami, kot jih uporablja klasično jezikoslovje.

Z upoštevanjem želja uporabnikov, nasvetov strokovnjakov, z uvajanjem novih znanstvenikov in raziskovalcev, usklajenimi normativnimi pravili ter pripravljenostjo za finančno podporo delu bi razvoj jezikovnih tehnologij pri nas vendarle lahko stekel tako, da se majhnost našega jezika ne bi več odražala tudi v tehnološkem zaostanku.

Jezikovne tehnologije obsegajo vrsto področij in značilnih (računalniških aplikacij), njihovo zgradbo, medsebojno povezanost in odvisnost pa si lahko ponazorimo na različne načine. Vsekakor je poleg razvoja informacijske tehnologije osnovni razlog za obstoj jezikovnih tehnologij obstoj naravnih jezikov samih, sistema pisave ter množica pisanih in govorjenih besedil, dostopni v raznih oblikah, predvsem elektronski.

9. LITERATURA

1. JURAFSKY, M.,H. MARTIN, J. *Speech and Language Prosessing*. str. 799-827
2. http://www.systransoft.com/products/index.html
3. http://cslu.cse.ogi.edu/HLTsurvey/HTLsurvey.html
4. http://nl.ijs.si/isjt04/zbornik/sdjt04-04holozan.pdf
5. BARTOLINI, B. Babilonski računalnik. *Moj mikro*, 2004, št.1, str. 64-65
6. ROMIH, M. Jezikovne tehnologije. *Življenje in tehnika*, januar 2003, str.20-29
7. ERJAVEC, T. Označevanje korpusov. *Jezik in slovstvo*, let. 48, št.3/4, str. 61-76
8. STABEJ, M. Jezikovne tehnologije in jezikovno načrtovanje. *Jezik in slovstvo*, let. 48, št.3/4, maj/avg 2003, str. 5-18
9. ERJAVEC, T. Jezikovne tehnologije za slovenski jezik. Mednarodna konferenca
10. HIRCI, N., PISANSKI, A. Nove jezikovne tehnologije: Vidiki uporabe računalniških korpusov. *Vestnik*, let. 34, št. 1/2 (2000), str. 27-34
11. http://www.fida.net/slo/index.html
12. PONIKVAR, M. *Računalniška podpora prevajalskemu in terminiloškemu delu na primeru prevajanja sektorja za prevajanje SVEZ* : diplomsko delo. Ljubljana, 2002