RAČUNALNIKI V PREVAJANJU: KAKO RAZVRSTITI PREDNOSTI?

Jaro Lajovic
Topniška 45, 1000 Ljubljana
Tel/faks: 061-1375-284
e-pošta: Jaro.Lajovic@mf.uni-lj.si

POVZETEK

Povečevanje informacijskega pretoka povečuje potrebo po prevajanju in računalniških orodjih zanj. Kljub medsebojni integraciji so ta orodja po zasnovi zelo različna, zato je njihovo zgradbo včasih smiselno obravnavati kot dva modula: programskega in jezikovnega. Pri razvrščanju razvojnih prioritet velja upoštevati, da so programski moduli sorazmerno jezikovno neodvisni, da je njihov razvoj dolgotrajen in da je bilo vanje vloženega že veliko dela drugod. Po drugi strani so jezikovni viri specifični in hkrati pomemben temelj za razvoj prevajalskih programov. Čeprav moramo slednjim (tudi razvojno) posvečati pozornost, je na področju jezikovnih tehnologij dolgoročno treba dati poudarek ustvarjanju jezikovnih zbirk: oblikovanju slovenskega nacionalnega korpusa, dvo- oz. večjezičnih korpusov s slovenščino in sorodnih baz (npr. splošnih in specialnih slovarjev).

ABSTRACT

Increased information flow increases needs for translation and, consequently, for the computerised translation tools (e. g. translation memories, machine translation programs). In spite of their integration, the concepts of these tools differ widely among different groups. All can, however, be regarded as consisting of a program module and a linguistic one. When setting the development priorities it should be considered that program modules are relatively language-independent, that their building is time-consuming and that much work has already been invested in them in several countries. On the other hand linguistic resources are language-specific and are at the same time an important basis for the development of translation tools. Although the former should not be neglected in nations with lesser-used languages (e. g. Slovenian), the long-term priority in the field of language technologies should be given to the latter, specifically to linguistic corpora. In Slovenia this would, inter alia, require the establishment of the Slovenian National Corpus, building of bi- and multilingual corpora including Slovenian, and creation of corresponding databases (e. g. general and specialised dictionaries).

1 UVOD

Zaradi različnih, tehnološko in politično pogojenih dogajanj (internet, globalizacija programske opreme, evropska integracija) se izrazito povečuje potreba po prevajanju [1], s tem pa tudi potreba in zanimanje za programsko opremo oz. orodja, ki lahko prevajanje olajšajo, pospešijo ali mu zagotovijo večjo kakovost. Gre za širok spekter pripomočkov, od najbolj znanih in uveljavljenih (npr. splošnih in specializiranih slovarjev, črkovalnikov) prek orodij za računalniško podprto prevajanje (zlasti prevodnih pomnilnikov - "translation memories"; TM) do računalniškega prevajanja ("machine translation"; MT).

Vsa ta orodja se danes marsikdaj tesno povezujejo, bodisi v okviru širših programskih rešitev (npr. integracija TM z MT) ali kompleksnejših sistemov, ki fungirajo kot celota (npr. SdT - prevajalska služba Evropske komisije)[2]. Poleg tega se zasnova nekaterih MT približuje zamisli TM (npr. na vzorcih temelječe računalniško prevajanje; EBMT). A čeprav je vsa ta oprema usmerjena v prevajanje, je med posameznimi skupinami več razlik kot podobnosti, deloma pa so namenjene tudi različnim skupinam uporabnikov. To je - celo znotraj skupine - zelo opazno pri računalniškem prevajanju, ki je lahko namenjeno pretežno asimilaciji ali diseminaciji informacij [3], kar (lahko) pomembno vpliva na zasnovo potrebnega oz. uporabljanega sistema ter potrebno kakovost prevoda.

2 PROGRAMI

Skrajno poenostavljeno lahko sestavo programov za računalniško in za računalniško podprto prevajanje obravnavamo kot dva modula: programskega in jezikovnega, "stroj" in "bazo znanja". V primeru (najbolj razširjenih "resnih") transferskih programov za računalniško prevajanje je programski modul sestavljen iz analitičnega (parser), transferskega in generativnega dela, jezikovni pa iz leksikona. V primeru TM je programski del sistem, ki omogoča delo z bazo in išče skladne segmente, jezikovni pa dvo- ali večjezični poravnani korpus. (V resnici strukture niso tako preproste; predvsem pri MT sta oba dela "organsko" prepletena, o pravi modularnosti ni govora in leksikoni niso prenosljivi. Obstajajo pa prizadevanja za oblikovanje standardov, ki

bi omogočili takšen prenos. Pri TM je ločitev modulov mnogo bolj stvarna in bližji so tudi standardi za prenos [4]). Podobno dihotomno delitev lahko dovolj smiselno uporabimo pri drugih orodjih.

Takšno dihotomno obravnavanje je uporabno, ker olajšuje razmislek o tovrstnih računalniških pripomočkih v zvezi s posameznim - v okviru tokratne razprave slovenskim - jezikom. Zanimanje za razvoj takšnih pripomočkov je odvisno predvsem od komercialnih in političnih potreb, možnosti za razvoj pa od znanja (znanstveno-raziskovalnih možnosti). A če o potencialu domačega znanja v tej smeri ni dvoma, si je bilo še pred kratkim težko predstavljati, da bi bila slovenščina komercialno ali politično dovolj zanimiva, da bi spodbudila zanimanje (in vlaganje) v takšen razvoj. Zdaj takšno zanimanje - v prvi vrsti politično - obstaja, eksplicitno izraženo s strani Evropske unije (zaradi iminentne pridružitve in s tem potrebnega dela, zlasti na področju zakonodaje) in ZDA (zaradi geopolitičnega položaja Slovenije)[5]. Ustanovitev Prevajalske službe za evropske zadeve pri slovenski vladi je verjetno znak podobnega zanimanja ali vsaj potrebe tudi doma.

Takšne okoliščine so priložnost, da dobimo ustrezna orodja za slovenski jezik - in pri tem izkoristimo lastno znanje, ne le uporabimo drugod razvite izdelke. Pritegniti velja vse, ki se s tem že ukvarjajo (npr. IJS, SAZU), potencialne domače uporabnike in - morda - tudi tuje partnerje. Seveda pa moramo ob tem pretehtati, kako razvrstiti prednosti. Ob kratkoročnih moramo razmisliti tudi o dolgoročnih - zlasti če se zavedamo, da so tisti sistemi za MT, ki pomenijo kakovost, plod dolgotrajnega (desetletnega) dela, in da tudi prevodni pomnilniki, ki so v primerjavi z njimi preprosti, niso nastali v letu ali dveh.

Če naj ne ponavljamo že prehojenega, velja prizadevanja usmeriti zlasti tja, kjer potrebnega dela ne more opraviti kdo drug (ali ga vsaj ne more opraviti enako kakovostno in/ali hitro). To pa so manj "programske" in bolj "jezikovne" komponente sistemov.

Ob tem se je dobro zavedati izkušenj evropskega projekta EUROTRA, da bi se izognili manj želenemu razvoju zadev. EUROTRA je bil projekt Evropske komisije, ki je potekal v letih 1982-92. Zastavljen je bil raziskovalno, vendarle tudi s ciljem, da bi ustvarili transferski MT sistem za prevajanje v Evropski komisiji. Sodelovali so raziskovalni (univerzitetni) centri držav članic in opravili precej raziskovalnega dela [6], ni pa to privedlo do praktičnega sistema (SdT uporablja Systranovega). Lingvistična usmeritev v komisiji se je nato spremenila in cilj LRE (Linguistic Research and Engineering), ki je sledil, je razvoj temeljne jezikoslovne tehnologije, ki jo je mogoče vključiti v različne aplikacije, ne samo v računalniško prevajanje [7].

Če ob tem upoštevamo, da več avtorjev kot priporočljivo pot za hitrejši razvoj programov za računalniško prevajanje svetuje avtomatično (ali polavtomatično) akvizicijo slovnice, pravil in drugega [8,9], se očitno ponuja misel, da lahko najbolj koristijo korpusi besedil. To pa pomeni, da mora biti prednostna naloga dolgoročnega razvoja računalniških orodij za slovenščino oblikovanje korpusov. Ti so interes tako znotraj "ožjega" področja (MT, TM, slovarji) kot širše v jezikoslovju [10], zato je najbrž mogoče pričakovati tudi dovolj široko podporo zanj. Kar zadeva računalniško prevajanje, bi korpusi lahko bili izhodišče za razvoj različnih sistemov, naj bodo direktni (kakršnega zasnove že obstajajo)[11], transferski, EBMT, statistični ...

3 OBLIKOVANJE KORPUSOV

Kako zagotoviti korpuse? Prvi, že zastavljeni korak v tej smeri je projekt FIDA [12]. Sicer pa se v odgovor (seveda ob ureditvi avtorskih pravic) ponujajo tri točke:

- Oblikovanje Slovenskega nacionalnega korpusa po zgledu takšnih že obstoječih. Izvedbeno to pomeni dopolnitev Zakona o knjižničarstvu tako, da bi osrednja nacionalna knjižnica dobivala obvezne izvode tudi na računalniškem mediju.
- Oblikovanje bi- ali multilingvalnih korpusov s slovenskimi besedili. To bo verjetno zahtevnejše, vendar bi lahko izhodišče pomenili prevodi pravnih besedil, ki bodo nastajali med pridruževanjem Evropski uniji. Potencialni viri so še drugi dokumenti, ki jih morajo npr. tuja podjetja zagotoviti v slovenščini, ali določeni založniški projekti.
- Nekoliko na meji konteksta korpusov, vendar vseeno pomembne v okviru jezikovnih tehnologij, so strokovne terminološke zbirke, glosarji in podobne baze. Pri njih je vprašanje avtorskih pravic (za neposredno uporabo) še bolj v ospredju, zato bi morali na nacionalni ravni v zvezi z njimi poskrbeti za dvoje. Podatki o takšnih zbirkah naj bi se stekali npr. v NUK in bili objavljeni na spletu ali bili dostopni prek COBISS, po drugi strani pa bi veljalo avtorje spodbuditi, naj jih pripravljajo upoštevaje uveljavljen format, ki bo omogočil njihovo integracijo v prevajalska orodja.

Nedvomno imajo jezikovne tehnologije - z računalniškim in računalniško podprtim prevajanjem kot pomembnima komponentama - veliko vlogo v informacijski družbi, saj olajšujejo pretok informacij in jih uporabniku zagotavljajo v njegovem jeziku. So lahko most zbliževanja in ekonomska spodbuda mnogim dejavnostim. Priložnosti se zagotovo ponujajo, mi pa se moramo glede njih odločiti za čim bolj plodne rešitve.

4 VIRI

1. Loughman L. The expanding world of globalization. Language International 1997; 9(2):21.

2. Computerised workflow in the Translation service. European Commission Translation Service. CDD (98)013; 25.02.1998.

3. Hutchins JW, Somers HL. An introduction to machine translation. London: Academic press, 1992.

4. http://www.lisa.org/tmx/index.html (Localisation Industry Standards Association: Translation Memory eXchange).
5. Pedtke TR. US government support and use of machine translation: current status. V: Teller V, Sundheim B, eds: Machine translation Summit VI proceedings. Washington: AMTA, 1997:3 ter osebna komunikacija.
6. Maegaard B. EUROTRA, history and results. V: MT Summit V proceedings. Luxembourg: EAMT, 1995.
7. Havenith R. MT research and development in Europe. V: Teller V, Sundheim B, eds: Machine translation Summit VI proceedings. Washington: AMTA, 1997:68.
8. Nirenburg S et al. Two principles and six techniques for rapid MT development. V: Expanding MT Horizons. Montreal: AMTA, 1996:96.
9. Hovy E. A gentle introduction to machine translation. Marina del Rey: Information Sciences Institute, 1996 (skripta).
10. Brekke M, Myking J, Ahmad K Terminology management and lesser-used living languages: a critique of the corpus-based approach. V: Galinski C, Schmitz KD, eds: Terminology and knowledge engineering. Frankfurt/Main: Indeks Verlag, 1996: 179.
11. http://www-ai.ijs.si/~ema/proj.html (Information System for Employment in Slovenia: a prototype of translating available jobs in English).
12. http://www.fida.net (FIDA: korpus slovenskega jezika).