Primož Jakopin

UDK 681.3.06 : 801.8

JEZIK, RAČUNALNIKI IN EVROPA OKOLI NAS

V prispevku je predstavljen sedanji trenutek na področju besedilnih podatkovnih zbirk, s posebnim poudarkom na projektih, ki v tej smeri za področje zahodne in vzhodne Evrope tečejo v okvirih raziskovalnih programov Evropske skupnosti.

A brief overview of the current action in the field of language corpora is given. Special emphasis has been given to the projects now under way or in preparation within the framework of the reseach programs, financed by the European Union.

Sredi septembra je bil na polotoku Tihany ob Blatnem jezeru na Madžarskem mednarodni seminar v okviru projekta TELRI in z zgovornim naslovom: Language Resources for Language Technology, tj. Jezikovni viri za jezikovno tehnologijo. Računalniki so bili seveda zelo v ospredju in če strojno prevajanje za vsakogar res še ni čisto vsakdanja stvar, pa se na področju računalniškega jezikoslovja vsaj zelo veliko dogaja.

Doslej je bilo seveda največ narejenega pri angleškem jeziku, ki ga ne proučujejo samo Angleži in Američani, ampak tudi veliko drugih raziskovalcev po svetu. Če ste, denimo, mlad raziskovalec, ki bi se rad uveljavil v svetu, je nekako najpreprosteje, če najprej napravite nekaj novega z angleškimi besedili. Tako na Nizozemskem, na Erazmovi univerzi v Nijmegnu, vzporedno gradijo tri velike besedilne zbirke (korpuse), povezane z ustreznimi računalniškimi slovarji, za angleški, nemški in holandski jezik. Največjo računalniško zbirko starih ameriških besedil pa imajo na Finskem.

Američani so se dela pri zbirkah lotili zelo načrtno in leta 1990 ustanovili konzorcij LDC (Linguistic Data Consortium), ki ga je izdatno podprla vladna agencija ARPA (American Research Projects Agency) in katerega osnovni namen je razpečevanje računalniških zbirk z govorjenimi in pisanimi jezikovnimi viri, predvsem za uporabnike iz gospodarstva. LDC ima tri vrste članov – prvi so velika podjetja, ki plačajo 50.000 dolarjev letne članarine in imajo pri vseh odločitvah glavno besedo, drugi majhna podjetja s članarino 2.000 dolarjev, akademske članice pa plačajo nekaj sto dolarjev letno. LDC je pri svojem delu tako uspešen, da so sklenili svojo dejavnost razširiti še na staro celino. S svojimi nameni so se obrnili kar naravnost na sedež Evropske skupnosti v Bruslju. Tam so eno leto razmišljali, potem pa je le prevladalo spoznanje, da organiziranje evropskih besedilnih zbirk v ameriški

režiji in pod ameriškimi pogoji ne bi pomenilo samo velikega ponižanja in sramote za skupnost, ampak bi bilo dolgoročno tudi manj smotrno. V marcu tega leta so zato sami ustanovili združenje ELRA (European Language Resources Association), katerega osnovna namena sta razpečevanje ocenjenih (validated) jezikovnih virov in pospeševanje njihovega širšega prodora. Letna članarina je 1.000 ECU-jev, članic pa je trenutno 30, od tega 15 akademskih in 15 iz gospodarstva. Imajo predvsem zbirke, ki se nanašajo na govor in manj takih s pisanimi besedili. Članstvo je bilo najprej omejeno na države Evropske skupnosti, Švico, Norveško in Lichtenstein, konec septembra naj bi ga razširili še na države srednje in vzhodne Evrope (znamenite dežele CEE). Piscu teh vrstic ob oddaji prispevka še ni bilo znano, pod kakšnimi pogoji.

Ustrezno evropsko združenje raziskovalnih ustanov, za dežele ES in na področju zbirk s pisanimi besedili se imenuje PAROLE (v francoščini beseda; WORD bi najbrž preveč spominjal na Microsoftov izdelek, pa še nič preveč evropsko ne zveni). Da vzhod, ki je strateško še kako pomemben, ne bi ostal čisto izven teh tokov, so se na Inštitutu za nemški jezik (IDS) v Mannheimu odločili, da poskusijo s projektom v okviru raziskovalnega programa Kopernik (Copernicus), ki je namenjen vzhodni Evropi in ga financira Evropska skupnost, povezati še te dežele. V bistvu se ponavlja zgodba o odnosu med Američani in Evropejci, le da vzhodna Evropa kake vidne lastne organizacije nima in se pusti iz zagat reševati tistemu, ki ga to pač zanima. Projekt ali bolje, združeno delovanje (Concerted Action) se imenuje TELRI, ki je kratica iz angleškega Trans-European Language Resources Infrastructure (Panevropska Infrastruktura Jezikovnih Virov), in bo trajal tri leta, tj. od začetka 1995 do konca 1997. Stal bo približno 200.000 ECU-jev, ki pa niso namenjeni raziskovanju, ampak samo sestankom družabnikov, organizaciji seminarjev, kratkim delovnim obiskom ipd. Sodeluje 22 akademskih ustanov, predvsem inštitutov pri fakultetah in akademijah znanosti, iz 17 držav: Albanije, Bolgarije, Češke, Estonije, Francije, Italije, Latvije, Litve, Madžarske, Nemčije, Nizozemske, Poljske, Romunije, Slovaške, Slovenije, Švedske in Velike Britanije. Dežele bivše Jugoslavije in Sovjetske zveze še niso vključene, so pa letos že navezali stike s Hrvaško in Srbijo ter Gruzijo - z Belorusijo, Ukrajino in Rusijo naj bi jih do konca leta.

Glavna pobusnika z zahodne strani sta že omenjeni Inštitut za nemški jezik iz Mannheima, kjer je sedež TELRI-ja in School of English iz Birminghama, kjer so na področju računalniškega jezikoslovja doslej še največ naredili. Slovenijo zastopata Inštitut za slovenski jezik Frana Ramovša ZRC SAZU (prof. dr. Varja Cvetko Orešnik ter podpisani) in Laboratorij za tehnologijo jezika in govora pri Inštitutu Jožef Stefan (mag.

T. Erjavec). Precej podatkov o TELRI-ju je mogoče izvedeti tudi prek Interneta, na WWW naslovu:
http://www.ids-mannheim.de/telri/telri.html ali neposredno po elektronski pošti: telri@ids-mannheim.de.

TELRI ima delo razdeljeno na 11 delovnih skupin (WorkGroups, koordinatorji so navedeni v oklepaju): uporabniška (W. Teubert, DE), dokumentacijska (R. Marcinkeviciene, LI), za pripravo biltena (E. Hajicova, CZ), za prirejanje seminarjev (J. Pajzs, HU), za pregled jezikoslovnega softvera v javni lasti (lingware, T. Erjavec, IJS, SI), za skupne usluge (P. Lafon, FR), za elektronsko mrežo (preko Interneta, V. Benko, SK), za povezovanje navzven (W. Teubert, DE), za skupne raziskave (J. Sinclair, GB), potrebe uporabnikov (A. Spektors, LA) in za trajno infrastrukturo (A. Zampolli, IT). Inštitut za slovenski jezik sodeluje v dveh skupinah: za skupne raziskave in za trajno infrastrukturo. Prva si je zadala za cilj primerjalno analizo enega literarnega dela, tj. Platonove Države, v kar največ jezikih iz sodelujočih držav. Platonova Država je bila izbrana zaradi svoje nevtralnosti, ponovne aktualnosti in zaradi številskih oznak pri odstavkih v originalnem besedilu, ki olajšuje poravnavo v različnih jezikih. Trenutno so zagotovljene elektronske verzije dela v enajstih jezikih, tudi v slovenščini. Predvidena je izdaja na dveh CD-jih, na enem maja 1996, ki bo vseboval poravnana (aligned) besedila, indeks besednih oblik in softver za iskanje, ter na drugem, septembra 1997, kjer bo besedilo opremljeno že tudi z oznakami (tags), indeksom besed in izborom prevajalnih ekvivalentov.

S podobno nalogo se ukvarja še en projekt v okviru programa Kopernik, tj. MULTEXT EAST, katerega naloga je zbrati primerljiva besedila, slovarje pojmov in ustrezni jezikoslovni softver, v različnih vzhodnih jezikih in tudi v slovenskem. Tam so si za primer dela v vseh jezikih izbrali znan Orwellov roman – 1984.

Udeleženci seminarja so lahko na povabljenih predavanjih in na predstavitvah skupnih projektov med akademskimi ustanovami in podjetji izvedeli veliko predvsem o gradnji, pomenu in izkoriščanju besedilnih zbirk ali korpusov, pa tudi o novostih na področju strojnega prevajanja in pri nekaterih drugih zelo aktualnih temah, npr. o razpoznavanju govora. Profesor Feng iz Pekinga je razložil, s kakšnimi problemi se srečujejo pri strojnem prepoznavanju kitajskih besedil (OCR), kjer npr. poleg izredno velikega števila pismenk, trenutno jih je 54.678, število pa se še povečuje, pojavljajo še problemi z določanjem stavčnih mej, ki niso posebej označene, kot pri nas s piko. V teku je desetletni projekt, s katerim nameravajo zbrati besedilno zbirko, dolgo 70 milijonov pismenk (približno toliko besed). Zanimivi so kriteriji za izbiro teh besedil:

1. diahronična omejitev: vsi viri morajo biti po letu 1919, poudarek pa je na gradivu po letu 1977 (po kulturni revoluciji);

2. kulturna omejitev: gradivo mora biti predvsem tako, da ga lahko razumejo ljudje s končano srednjo šolo;

3. omejitev rabe: gradivo mora biti iz splošne rabe, pri čemer morajo imeti prednost družboslovne znanosti in humanistika.

Videli smo lahko tudi, koliko zaostajamo za bolj razvitimi sosedi, npr. za Angleži. Besedilne zbirke se zdaj ne merijo več s KW (niso Kilovati ampak KiloWords = tisoč besed) ampak z MW (MegaWords = milijon besed). Da neki zbirki z besedili lahko rečete korpus, mora biti vsebinsko in namensko zaokrožena ter urejena po standardih. Spodobi se tudi, da ni majhna: angleška zbirka, ki jo najbolj uporabljajo in ki je na univerzi v Birminghamu, se imenuje Bank of English in obsega 200 MW. Če so bili njihovi korpusi v obdobju od 1965-75 veliki po 1 MW, od 1975-85 po 20 MW, od 1985-95 po 200 MW, lahko nas s po približno 5 MW dolgimi besedilnimi zbirkami postavimo nekam v leto 1980. Za primerjavo povejmo, da ima 350 strani dolga knjiga približno sto tisoč besed.

Tudi projekti na več jezikih vzporedno so v načrtu. Najpomembnejši je CORDON (Corpus-ORiented Detection Of Neologisms), ki bo trajal dve leti (1996-97), združeval po štiri akademske partnerje in po štiri iz gospodarstva za angleški, nemški, francoski in švedski jezik. Akademski del posla, v obsegu 96 raziskovalnih mesecev, bo stal 1.7 milijona ECU-jev, ki jih bo dala Evropska skupnost. Namen projekta, da je s pomočjo primerljivih besedilnih zbirk s po 50 milijoni besed razvije modularno, od jezika neodvisno programsko opremo, ki bo znala poiskati nove besede in fraze, ki označujejo nove pojme. Z njo bi v prihodnje lažje zagotovili aktualnost jezikoslovnih virov, orodij in terminoloških podatkovnih zbirk.

Pokažimo še, kako zapleteno bo življenje v prihodnjem tisočletju, in sicer na primeru novih standardov za računalniško shranjevanje in izmenjavo besedil. Nancy Ide, predstavnica konzorcija TEI, Pobude za kodiranje besedil (Text Encoding Initiative), je predstavila na kratko, kako se je treba lotiti knjige, slovarja, drame ali česarkoli drugega pisanega, da jo bodo poleg vašega razumeli še drugi programi za jezikoslovne obdelave. Projekt so zastavili leta 1987, v okviru treh društev: Association for Computational Linguistics, Association for Computers and the Humanities in Association for Literary and Linguistic Computing. Prvi dve sta ameriški, zadnje pa je evropsko. Bilo je veliko prostovoljnega dela, glavnino denarja so prispevali: U.S. National Endowment for the Humanities, Commission of the European Community DG XIII ter Mellon Foundation. Kar so naredili, je v knjižni obliki in na CD-ju izšlo leta 1994: Guidelines for Electronic Text Encoding and

Interchange (TEI P3), Dynatext Edition. Naročiti se da CD za 50 funtov na naslovu: TEI Orders, Oxford University Computing Services, 13 Banbury Road, Oxford OX2 6NN, Velika Britanija. Naročilo je možno poslati tudi prek elektronske pošte. Dovolj je, da na elektronski naslov pri univerzi v Chicagu:

listserv@uicvm.uic.edu

pošljete eno izmed naslednjih vrstic:

get p3ascii package
get teip3 package

get p3dtds package
get p3all package

S prvo zahtevate in v nekaj minutah tudi dobite TEIjeve smernice (1.300 strani) v običajnem ASCII formatu, z drugo v obliki, ki vsebuje tudi SGML-jeve oznake, s tretjo le opise dokumentov (angl. DTDs = Document Type Definitions), s četrto pa vse tri navedene pošiljke hkrati.

Osnovna standarda pri TEI sta ISO 8879, ki določa jezik za označevanje SGML (Standard Generalized Markup Language), ter ISO 646, ki določa sedembitni nabor znakov s katerim je opisano še vse ostalo. Besedo "taščica" bi po SGML recimo zapisali kot "ta&chacek;&shacek;ica". Da bo stvar še malo jasnejša (ali pa zapletenejša), poglejmo, kako bi po navodilih TEI zapisali angleško slovarsko geslo za besedo zapustiti. Najprej jo navedimo v klasični slovarski obliki –

a.ban.don 1 /@"b&nd@n/ v [T1] 1 to leave completely and for ever; desert: The sailors abandoned the burning ship. 2 ... abandon 2 n [U] the state when one's feelings and actions are uncontrolled; freedom from control: The people were so excited that they jumped and shouted with abandon / in gay abandon. [LDOCE]

kjer je z LDOCE označen vir (Longman Dictionary of Contemporary English), potem pa še v skladu z navodili TEI:

```
<superEntry>
  <form>
    <orth>abandon</orth>
    <hyph>a|ban|don</hyph>
```

```
      <pron>@"b&amp;amp;nd@n</pron>
    </form>
    <entry n='1'>
      <gramGrp> <pos>v</pos> <subc>T1</subc> </gramGrp>
      <sense n='1'><def>to leave completely and for ever
...</def>
        <!-- ... -->

      </sense>
      <sense n='2'>
        <!-- ... -->
      </sense>
    </entry>
    <entry n='2'>
      <gramGrp> <pos>n/pos> <subc>U</subc> </gramGrp>
      <def>the state when one's feelings and actions are
            uncontrolled; freedom from control</def>
      <!-- ... -->
    </entry>
  </superEntry>
```

Po TEI bomo najbrž vsi morali. Si pa lahko predstavljate, koliko dela, kljub računalniškim bližnjicam, bo, da spravimo npr. Slovar slovenskega knjižnega jezika s 93.151 gesli v tako obliko?

Novost s seminarja, zanimiva tudi za marsikoga izmed nas, je nov časopis, International Journal of Corpus Linguistics (IJCL), ki bo začel izhajati v začetku 1996, predvidoma štirikrat letno in ki je namenjen predvsem problemom, povezanim z obdelavo računalniških besedilnih zbirk in raziskavami sporočil v običajnem jeziku (NLP – Natural Language Processing). Naslov za poizvedbe je IJCL, John Benjamins Publishing Company, P.O. Box 75577, 1070 AN Amsterdam, Nizozemska.

Ob koncu lahko rečemo, da bo treba v prihodnje, če želimo, da slovenščina ostane jezik za vsestransko uporabo tudi še po letu 2000, veliko delati, tudi ob računalnikih.

VIRI IN LITERATURA

E. van Herwijnen, *Practical SGML.* – Kluwer, Amsterdam, 1990

N. Ide, J. Veronis, *Encoding Print Dictionaries.* – Posebna izdaja revije Computers and the Humanities, 1994

Text Encoding Initiative. – Guidelines for Electronic Text Encoding and Interchange (TEI P3, 1994). Naročila: TEI Orders, Oxford University Computing Services, 13 Banbury Road, Oxford OX2 6NN

TELRI – Trans European Language Resources Infrastructure. – Concerted Action in the Framework of the Copernicus Program, Newsletter 1, Praga, September 1995.

Summary

In the paper the events, related to the seminar, organized in the Hungarian resort of Tihany in September 1995 by TELRI, Trans European Language Resources Infrastructure project, are described. TELRI, a concerted action in the framework of the Copernicus program, financed by European Union and led by the Institut fuer deutsche Sprache (IDS) in Mannheim, has brought together 22 academic partners from 17 countries, mostly from the so-called Central and Eastern Europe (CEE countries) but also from United Kingdom, Germany, France, The Netherlands, Sweden and Italy.

Large balanced corpora of text, parsed and annotated, in the region of 200 milllion words and over, are now regularly updated and maintained for the mainstream languages of quantitative linguistics research. The most notable example is the corpus Bank of English at the University of Birmingham in the UK. The research itself is now moving from the corpus-referenced phase with strong hypothesis, constructed by the researcher's impressions about language, which are verified on the textual data base, to corpus-driven research, where large number of weak hypothesis, coming from the work on the corpus, are taken into consideration.

The efforts in the CEE countries are more fragmented and have less tradition; to preserve their culture and national heritage this part of the world is now, within frames of the possible and with open insight of what is going on in the West, trying to catch up.