KORPUS FIDA

Tomaž Erjavec,[1] Vojko Gorjanc,[2] Marko Stabej[2]

(1) Odsek za inteligentne sisteme, Institut Jožef Stefan
Jamova 39, Ljubljana
tomaz.erjavec@ijs.si

(2) Oddelek za slovanske jezike in književnosti, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, Ljubljana
vojko.gorjanc@guest.arnes.si, marko.stabej@guest.arnes.si

POVZETEK

V okviru projekta FIDA je v izdelavi referenčni korpus slovenskega jezika. Članek predstavi projekt in opiše zvrstnost ter zgradbo korpusa FIDA. Posebej opiše način vključevanja besedil v korpus in digitalni zapis korpusa.

ABSTRACT

The FIDA project is compiling a reference corpus of the Slovene language. The paper introduces the project nad describes the characteristics and structure of the FIDA corpus. The methodology of incorporating texts into the corpus and the digital coding of the corpus is also discussed.

1 UVOD

FIDA je projekt, ki ga skupaj pripravljajo pogodbeni partnerji Filozofska fakulteta Univerze v Ljubljani, Institut Jožef Stefan, založba DZS, d. d. in podjetje Amebis, d. o. o. Cilj projekta je oblikovanje referenčnega elektronskega korpusa besedil slovenskega jezika. Prva faza projekta traja od 1. 1. 1997 do 1. 7. 1999.

Že nekaj let se je na marsikaterem področju dejavnosti, povezanih z raziskovanjem in opisovanjem slovenskega jezika, vse očitneje kazala potreba po dovolj obsežnem, reprezentativnem in dostopnem korpusu, ki bi zagotavljal objektiviziran pogled na jezik in omogočal uporabo sodobne računalniške tehnologije tako pri temeljnih jezikoslovnih in drugih raziskavah kot pri razvijanju najrazličnejših programskih orodij za obdelavo besedil, predvsem v tistih delih, kjer morajo biti prilagojeni posameznemu naravnemu jeziku.

S projektom FIDA se pravzaprav nadaljujejo in nadgrajujejo prejšnje oblike sodelovanja med omenjenimi partnerji. Po eni strani se je namreč izkazalo, da velik del sodobnega slovenskega teoretičnega in uporabnega jezikoslovja, npr. besediloslovje in leksikografija, svojih nalog brez korpusa preprosto ne more več opravljati, saj ne izpolnjuje več niti najnižjih mednarodnih meril, ki so se temeljito preoblikovala prav zaradi korpusnega pristopa. Po drugi strani pa je postalo očitno, da zaradi kompleksnosti naloge izdelava korpusa ne more biti stvar enega samega partnerja, temveč nujno zahteva tesno sodelovanje med strokami in ustanovami. Poleg tega projekt združuje, kar zlasti za humanistični prostor v Sloveniji ni tako pogosto, znanstvenoraziskovalni ustanovi in kapitalski družbi. Raziskovalci tako niso le najeta honorarna delovna sila, ampak s pogodbenim razmerjem svojim ustanovam in širši raziskovalni javnosti jamčijo dostop do raziskovalnih rezultatov. Kapitalske družbe pa niso le pasivni založniki ali uporabniki nekega raziskovalnega dosežka, temveč s svojim vlaganjem raziskovalno delo omogočajo in ga seveda – z zahtevo po terminsko natančno določenih stopenjskih rezultatih – tudi nadzorujejo. Projekt torej že od samega začetka združuje partnerje z različnim znanjem in potrebami, kar se zaenkrat kaže kot učinkovitejša pot od tistih projektov v povezavi z jezikovnimi tehnologijami, ki so izhodiščno zasnovani samo v enem institucionalnem okviru in šele naknadno iščejo širše povezave, pri čemer je marsikdaj onemogočeno revidiranje izhodiščnih hipotez, čeprav se to izkaže ob poznejšem sodelovanju za nujno potrebno.

2 KARAKTERISTIKE KORPUSA

Korpus FIDA je *enojezikovni korpus*, vanj bodo vključena sodobna slovenska besedila, kar pomeni, da se tujejezični elementi v korpusu lahko pojavijo le kot sestavni del slovenskega besedila, izključena pa so vsa tujejezična, npr. tudi italijanska iz dvojezikovnih medijev na Obali. Hkrati je *sinhroni korpus*, torej korpus sodobne slovenščine druge polovice 20. stoletja, vendar s poudarkom na zajemanju besedil, nastalih v zadnjih dvajsetih letih.

FIDA bo kot *referenčni korpus* oblikovan tako, da bo lahko posredoval vsestranske izčrpne informacije o slovenskem jeziku. Da bi bile informacije relevantne, je potrebno zagotoviti dovolj veliko količino raznovrstnih besedil, tako da korpus predstavlja uravnoteženo reprezentativno elektronsko besedilno zbirko. Prav vprašanji uravnoteženosti in reprezentativnosti sta pri postavitvi referenčnega korpusa ključni. Za zagotavljanje uravnoteženosti in s tem tudi reprezentativnosti je v izhodišču oblikovana mreža parametrov, s pomočjo katerih se določa količina vključevanja različnih besedil v korpus. Za lažje zagotavljanje uravnoteženosti je celotni korpus notranje hierarhiziran v podkorpuse, npr. časopisnih besedil, ki imajo lahko različne komponente, npr. Delo.

Vsaj v svojem izhodišču je FIDA *pisni korpus*; zajema torej pisna besedila in prvotno pisna besedila, namenjena govorjenju. Vendar se v zalogi besedil FIDA shranjujejo tudi transkribcije govora, npr. parlamentarne razprave, z ambicijo oblikovati tudi podkorpus govornih besedil.

Po tipologiji Eagles so karakteristike nekega korpusa tudi velikost, kakovost, avtentičnost in dokumentiranost [5]. Tako je eden od parametrov doseganja reprezentativnosti *velikost korpusa*; ta naj bi s svojim obsegom zagotovil dovolj veliko količino jezikovnih podatkov. Velikost korpusov danes narašča zelo hitro, saj je ob vzpostavljeni dinamiki pritoka besedil v korpus vse lažje zagotoviti veliko količino besedil. Ker so korpusi dinamična pojavnost, je tudi pri korpusu FIDA določeno le izhodišče, tj. 100 milijonov besed kot merilo reprezentativnosti [2, 10], vendar z idejo nadaljnjega sprotnega spremljanja jezikovnega dogajanja s stalnim vključevanjem novega besedilnega gradiva.

Kakovost vsakega korpusa v izhodišču določa *avtentičnost besedil* [5]. Korpusi naj bi predstavljali jezik v realni rabi, tako da v korpusu lahko pričakujemo le avtentična besedila. Ker je pomembno merilo avtentičnosti tudi morebitni jezikovni poseg v avtorsko besedilo [5], v slovenskem primeru npr. lektorski, je v glavo dokumenta, če ta podatek obstaja, vključen tudi podatek o lektoriranju.

Dokumentiranost je zagotovljena s podatki o besedilu v glavi vsakega dokumenta, vključenega v korpus. Poleg obveznih podatkov v glavi dokumenta je v primeru, ko že obstaja kataloški zapis v sistemu COBISS, tudi ta vključen v glavo posameznega dokumenta.

Vsako besedilo, vključeno v korpus FIDA pa je opredeljeno tudi glede na taksonomijo, definirano v glavi celotnega korpusa FIDA. Taksonomija z identifikatorji in opisnimi imeni je podane v tabeli 1.

Ft Taksonomija FIDA
- **Ft.P prenosnik**
  - Ft.P.G govorni
  - Ft.P.E elektronski
  - Ft.P.P pisni
    - Ft.P.P.O objavljeno
      - Ft.P.P.O.K knjižno
      - Ft.P.P.O.P periodično
        - Ft.P.P.O.P.C časopisno
          - Ft.P.P.O.P.C.D dnevno
          - Ft.P.P.O.P.C.V večkrat tedensko
      - Ft.P.P.O.P.R revialno
        - Ft.P.P.O.P.R.T dvomesečno
        - Ft.P.P.O.P.R.M mesečno
        - Ft.P.P.O.P.R.D redkeje kot na mesec
    - Ft.P.P.N neobjavljeno
      - Ft.P.P.N.J javno
      - Ft.P.P.N.I interno
      - Ft.P.P.N.Z zasebno
- **Ft.Z zvrst**
  - Ft.Z.U umetnostna
    - Ft.Z.U.P pesniška
    - Ft.Z.U.R prozna
    - Ft.Z.U.D dramska
  - Ft.Z.N neumetnostna
    - Ft.Z.N.S strokovna
      - Ft.Z.N.S.H humanistična
      - Ft.Z.N.S.D družboslovna
      - Ft.Z.N.S.N naravoslovna
      - Ft.Z.N.S.T tehnična
    - Ft.Z.N.N nestrokovna
    - Ft.Z.N.P pravna
- **Ft.L lektorirano**
  - Ft.L.D da
  - Ft.L.N ne

Tabela 1: Taksonomija FIDA

3 VKLJUČEVANJE BESEDIL V KORPUS

V zalogi besedil FIDA se zbirajo in shranjuje elektronska besedila v izvirni obliki, torej taka, kot jih za korpus pridobimo od različnih besedilodajalcev.

Z vsakim besedilodajalcem je podpisana enotna pogodba o odstopu besedil v elektronski obliki, ki zagotavlja avtorjem oz. imetnikom avtorske pravice nad besedilom vse pravice, hkrati pa ureja načine, na katere lahko projekt razpolaga z besedili za vse potrebne nadaljnje obdelave in formatiranja.

Načela zajemanja v zalogo besedil FIDA so deloma prekrivna z načeli drugih korpusov podobnega obsega, deloma pa so bila oblikovana tudi dodatna načela, npr. načelo regionalne uravnoteženosti pri tiskanih medijih. Samo zajemanje besedil ni količinsko restriktivno, saj je cilj zbrati čim več besedil (kar je v slovenskem prostoru zaradi še ne dovolj razvite kulture elektronskega shranjevanja in arhiviranja besedil težje, kot je bilo na začetku projekta pričakovano).

Iz zaloge besedil se s postopki elektronske obdelave pridobi enoten format besedil, ki šele standardizirano zapisana tvorijo korpus FIDA. Kriteriji uvrščanja besedil v korpus pa so oblikovani tako, da korpus FIDA predstavlja uravnoteženo reprezentativno besedilno zbirko.

Hkrati z izgradnjo korpusa FIDA se oblikujejo merila njegove uravnoteženosti. Oblikovana je strategija postopnega vzpostavljanja mreže kriterijev za doseganje uravnoteženosti in s tem reprezentativnosti. Reprezentativnost je sicer relativna kategorija, saj je nemogoče predvideti in v korpus zajeti vse besedilne variante, vendar pa se skuša z merili reprezentativnosti zajeti vsaj ključne, ki pa morajo vključevati čim več jezikovnih variant.

Količinska razmerja med različnimi besedili so v izhodišču odvisna predvsem od recepcije različnih besedil, pa tudi besedilne produkcije [1, 4]. Predvsem z vidika recepcije se za doseganje reprezentativnosti oblikujejo parametri glede na jezikovno zvrst, besedilno vrsto, žanrsko pripadnost, medij, v katerem se pojavljajo, ipd. Za določanje razmerij med besedili posameznih jezikovnih zvrsti in besedilnih vrst se upošteva razpoložljive podatke o branosti, npr. podatki

Mediane, pa tudi ankete, oblikovane posebej za ta namen; ankete branosti so kriterij npr. tudi pri določanju količine besedil posameznih literarnih zvrsti v okviru umetnostnih besedil (poezije, proze, dramatike), pri tem pa so upoštevani tudi podatki o knjižnični izposoji. Nadalje so upoštevani npr. tudi Medianini podatki o nakladi, nadgrajeni s podatki o prostoru, ki ga posamezni medij pokriva, dosegu, ciljni skupini ipd.
Merila reprezentativnosti upoštevajo tuje izkušnje, vendar se za slovenščino glede na specifike našega prostora v veliki meri oblikujejo popolnoma na novo, zato je razumljivo, da se ob srečevanju s konkretnimi problemi in v diskusiji v okviru korpusne skupine (vse bolj pa tudi širše) dinamično prilagajajo.

4 ZAPIS KORPUSA

Ker imajo računalniški korpusi mnogotere uporabe, v njihovo izdelavo pa je potrebno vložiti precejšnjo količino dela, je smiselno zagotoviti njihovo čim večjo izmenljivost in jih zavarovati pred zastaranjem. To dosežemo z upoštevanjem mednarodnih standardov in priporočil pri zapisu korpusa. Standardni zapisi so namreč natančno dokumentirani, javno dostopni in neodvisni od specifičnega programskega okolja. Korpus FIDA bo zato zapisan po priporočilih TEI (Text Encoding Initiative), ki so aplikacija ISO-standarda SGML (Standard Generalised Markup Language). Priporočila TEI [8], t. i. TEI P3, se dandanes mednarodno uporabljajo pri zapisu večine računalniških korpusov. Uporaba TEI ima prednost v mednarodni primerljivosti zapisa, njena slabost za (pretežno) slovenski korpus pa je, da izhajajo imena oznak iz angleškega jezika. FIDA rešuje ta problem s prevajanjem oznak v opisne izraze v slovenščini.
TEI P3 podajajo definicijo oznak, ki služijo za opis široke palete jezikovnih zvrsti (npr. proza, slovarji) in interpretacij besedil (uredniška, literarnokritična, jezikoslovna). Da iz TEI P3 dobimo specifično definicijo tipa dokumentov, ga je potrebno parametrizirati, tako da ustreza specifičnosti besedil in interpretaciji le-teh oz. potrebam projekta, v okviru katerega naj bi besedila označevali.
Začetna parametrizacija TEI P3 za potrebe korpusa FIDA je podobna definiciji tipa dokumentov TEIlite ('lahki TEI', [9]). Definicija tipa FIDA zajema štiri module TEI: TEI.prose (leposlovje), TEI.linking (navzkrižne povezave), TEI.analysis (osnovni elementi jezikovne analize) in TEI.figures (slikovni elementi). Nadalje vsebuje definicija tipa dokumenta FIDA še parametrizacijo posameznih elementov TEI, predvsem tistih iz glave besedila. Tu so nam bila za vodilo priporočila CES (Corpus Encoding Specification), ki smo jih pred tem uporabljali za slovenski korpus v projektu MULTEXT-East [7].

```
<teiCorpus.2 lang="sl">        <!--Korpus FIDA-->
  <teiHeader type="corpus">  <!--Glava korpusa-->
    <fileDesc>…</fileDesc>
    <encodingDesc>…</encodingDesc>
    <profileDesc>…</profileDesc>
    <revisionDesc>…</revisionDesc>
  </teiHeader>
  <tei.2>                          <!--Besedilo 001-->
    <teiHeader type="text">
     ...                      <!--Glava besedila 001-->
    </teiHeader>
    <text><body>
     ...                       <!--Telo besedila 001-->
    </body></text>
  </tei.2>
  <tei.2>…</tei.2>               <!--Besedilo 002-->
  …
  <tei.2>…</tei.2>               <!--Besedilo 999-->
</teiCorpus.2>
```

Tabela 2: Vrhnje oznake korpusa FIDA

Definicija tipa dokumenta FIDA določa oznake, ki se nato uporabljajo pri zapisu korpusa. Shematska zgradba korpusa je podana v tabeli 2.Korpus kot celota, pa tudi vsako posamezno besedilo znotraj korpusa ima svojo glavo, tj. element z oznako <teiHeader>. Ta vsebuje informacije o korpusu kot celoti oz. o konkretnem besedilu ter služi kot dokumentacija korpusu oz. besedilu. Glava TEI oz. FIDA vsebuje štiri vrhnje elemente. V opisu datoteke z oznako <fileDesc> najdemo podatke o korpusu oz. besedilu, vključno z bibliografskimi podatki in navedbo virov. V opisu zapisa <encodingDesc> so opisane oznake v korpusu oz. besedilu; tu je podano, katere in kolikokrat se posamezne oznake v besedilu uporabijo in v primeru, da so dodatno definirane glede na nadrejeno dokumentacijo, kaj pomenijo. V opisu zapisa glave celotnega korpusa je tudi definirana taksonomija korpusa FIDA. Opis profila <profileDesc> podaja nebibliografske podatke o besedilu. V glavi korpusa FIDA so definirani identifikatorji jezikov, v glavi besedil pa, kam spada besedilo v taksonomiji korpusa FIDA. Končno vsebuje glava še opis sprememb <revisionDesc>, ki beleži spremembe v korpusu oz. besedilu.
Celoten korpus FIDA je sestavljen iz glave in besedil. Vsako besedilo ima spet svojo glavo ter telo <body>, v katerem je zapisano besedilo samo. Telesa korpusa so pri pretvorbi iz originalnega zapisa besedila očiščena originalnih (npr. RTF) oznak, neASCII znaki pa bodo zapisani kot standardne entitete SGML (npr. 'č' za č). Pri pretvorbi iz originalnega digitalnega zapisa bodo v telesih zabeležene naslednje oznake: odstavek, <p>; poudarjeno, <hi> (ta z atributom pove tudi to, kako je element poudarjen); veren zapis, <orig> (kjer zaradi neznanih znakov ni mogoče pretvoriti originala v zapis FIDA); manjkajoče, <gap> (kjer del originala ni bil zajet).
Bolj zanimive oznake se bodo pojavile v zadnji fazi projekta, ko je na vrsti jezikoslovno označevanje. Tu se bo označilo povedi, <s>, in besede, <w>, ki pa skozi

vrednost atributa lahko zajemajo tudi oblikoslovne značilnosti besede v besedilu.

5 ZAKLJUČEK

FIDA je prvi večji korpusni projekt v Sloveniji. S tem se pridružujemo jezikoslovnim težnjam v svetu, ki so od konca 60-ih let temeljito spremenile pogled na jezik. Veliko evropskih jezikov je že zajetih v besedilne korpuse in tuje izkušnje jasno izpričujejo, kako spodbudno to vpliva na razvoj uporabnega jezikoslovja, npr. slovaropisja v vseh oblikah (sodobnejši in dostopnejši eno- in več- jezikovni slovarji, terminološki slovarji in drugi jezikovni priročniki), poučevanja jezika (učbeniki in učni pripomočki) in jezikovne tehnologije (črkovalniki, slovnični pregledovalniki, govorni vmesniki), hkrati pa tudi na razvoj teoretičnih pogledov na jezik in komunikacijo. Vse to pa pomeni pomemben korak k večjemu jezikovnemu znanju in splošni jezikovni kulturi.

ZAHVALA

Trenutno so v zalogo besedil FIDA vključene publikacije *DZS* in *Znanstvenega inštituta Filozofske fakultete,* stalno pa pritekajo besedila dnevnika *Delo*. Med lokalnimi tiskanimi mediji je dogovor sklenjen z *Dolenjskim listom*, *Notranjskimi noticami*, *Novim tednikom Celje*, *Portorožanom*, *Primorskimi novicami* in *Panoramo Slovenska Bistrica*, v zalogi besedil pa so tudi revije *Annales*, *Acta Histria*, *Glamur*, *Lipov list*, *Moj pes*, *Muska* in *Tim.* Vsem naštetim se zahvaljujemo za besedila, ki so jih dali na razpolago projektu FIDA.

6 LITERATURA

[1] Douglas BIBER, Susan CONRAD, Randi REPPEN, 1998: Corpus Linguistics. Investigating language Structure and Use. Cambridge: Cambridge University Press.

[2] BNC – British National Corpus. http://info.ox.ac.uk/bnc/

[3] František ČERMÁK, 1995: Jazikový korpus: Prostredek a zdroj poznání. Slovo a slovesnost 56. 119–140.

[4] František ČERMÁK, 1997: Czech Natonal Corpus: A Case in Many Contexts. International Journal of Corpus Linguistics 2/2. 181–197.

[5] EAGLES: Expert Advisory Group on Language Engineering Standards. http://www.ilc.pi.cnr.it/EAGLES96/home.html

[6] Tomaž ERJAVEC, 1996/97: Računalniške zbirke besedil. Jezik in Slovstvo 2-3/42. 81–96. http://nl.ijs.si/tomaz/Bib/SlKorpus/slKorpus-la2l

[7] Tomaž ERJAVEC, Nancy IDE, 1998: The MULTEXT-East Corpus. First International Conference on Language Resources and Evaluation, LREC'98. Ur. Antonio Rubio, Natividad Gallardo, Rosa Castro, Antonio Tejada. Granada. 971–974. http://nl.ijs.si/ME/

[8] TEI94 – Guidelines for Electronic Text Encoding and Interchange, 1994. Ur. C. M. Sperberg-McQueen, Lou Burnard. Chicago, Oxford. http://www-tei.uic.edu/orgs/tei/

[9] Lou BURNARD, C.M. SPERBERG-MCQUEEN, 1995: TEI Lite: An Introduction to Text Encoding for Interchange**.** http://www.uic.edu/orgs/tei/lite/

[10] ICNC – The Institute of the Czech National Corpus. http://ucnk.ff.cuni.cz