AMEBIS IN JEZIKOVNE TEHNOLOGIJE

Miro Romih
Amebis d. o. o.
p. p. 69, 1241 Kamnik
miro.romih@amebis.si

POVZETEK

Podjetje Amebis d.o.o. se z jezikovnimi tehnologijami ukvarja že od leta 1990, pri čemer še posebno pozornost posveča slovenskemu jeziku. Članek opisuje aktivnosti podjetja Amebis d.o.o. na tem področju znanosti. Posebej so izpostavljeni črkovalnik, delilnik, besedna analiza, tezaver, korpus, elektronski slovarji, strojno podprto prevajanje in sinteza govora.

ABSTRACT

The company Amebis d.o.o. has been involved in language technologies since the beginning of the 1990's, concentrating in particular on the Slovene language. The paper introduces the activities of Amebis d. o. o., with special attention given to the areas of spell-checking, hyphenation, syntax checking, thesaurus, corpus, electronic dictionaries, machine translation and speech synthesis.

1 UVOD

Prispevek opisuje stanje na področju jezikovnih tehnologij v podjetju Amebis d.o.o. Desetletje že mineva, kar smo pričeli z razvojem na tem vse bolj pomembnem področju. Rezultat tega razvoja so različni uporabni programi in jezikovni moduli, ki jih tudi veliki svetovni proizvajalci programske opreme vgrajujejo v svoje proizvode. Sem sodita predvsem črkovalnik in delilnik, na pravo uporabnost pa čakata še tezaver in slovnični modul.

Za razvoj omenjenih jezikovnih modulov je poleg ustreznega računalniškega in jezikovnega znanja potrebna tudi obsežna baza besedil, iz katerih črpamo podatke in jih ustrezno predelamo. Zato je pomemben del našega delovanja posvečen tudi skrbi za gradnjo korpusa slovenskih besedil, ki se jih je v teh letih nabralo že kar precej.

Zelo pomembno področje jezikovnih tehnologij so tudi različni enojezični in dvojezični elektronski slovarji, ki iz enostavne računalniške izvedbe knjižnih izdaj počasi preraščajo v prava orodja za strojno (podprto) prevajanje.

Tudi področje razpoznavanja in sinteze govora vse bolj dobiva na veljavi. Ker si je množico sodobnih naprav in sistemov praktično nemogoče zamišljati brez teh tehnologij, del energije vlagamo tudi na ta področja.

2 ČRKOVALNIK

Osnovno jezikovno orodje nekega jezika je črkovalnik. To je program, katerega funkcija je odkrivanje tipkarskih napak v besedilu. Sestavljen je iz programskega in slovarskega dela.

Programski del je izveden glede na okolje, v katerem mora delovati. Njegovi osnovni funkciji sta preverjanje, ali beseda v slovarju obstaja ali ne, in dajanje nasvetov. Preverjanje je v osnovi relativno enostavna stvar, saj mora program le pogledati, ali ima v slovarju ustrezno besedno obliko. Če je slovar abecedno urejen, kar je običajno, je preverjanje zelo hitra operacija. Dajanje nasvetov je nekoliko bolj zahtevno opravilo, saj bi moral program za dano besedo teoretično pregledati celoten slovar in poiskati po obliki najbližje besedne oblike. Ker tako preverjanje zahteva preveč dragocenega časa, si je potrebno pomagati z določenimi algoritmi in triki, ki iskanje pohitrijo, predvsem pa omogočijo kvalitetne zadetke. Eden takih trikov je npr. vgrajen algoritem, ki na osnovi verjetnostne tabele, zgrajene s pomočjo predhodne statistične analize slovenskega jezika, omogoča pohitritev iskanja z izločanjem neverjetnih kombinacij črk.

Slovarski del črkovalnika je abecedno urejen spisek besed, ki ga programski del uporablja pri svojem delu. Zaradi hitrejšega iskanja po slovarju je največkrat potrebno tudi dodatno indeksiranje, v komercialnih programih pa je dodana še zahteva po čim manjši porabi prostora, zato je potrebno slovar še dodatno stiskati.

Seveda pa je pri vsem najpomembnejša vsebina oz. število in kvaliteta vsebovanih besednih oblik, ki jih slovar vsebuje. Naš črkovalnik trenutno vsebuje preko 550.000 različnih preverjenih besednih oblik, izvedenih iz več kot 42.000 najpogostejših osnovnih besed. Slovar gradimo že od leta 1988 in ga stalno dopolnjujemo. Dnevno dodajamo nove besede oz. besedne oblike. Zanimivo je, da je bolj kot količina vsebovanih besed pomembna njihova izbira. Tako lahko z nekajkrat manjšim številom pravilno izbranih besed dosežemo celo boljše rezultate preverjanja kot s slovarjem, ki vsebuje precej več besed, ki pa niso selektivno dodane. Frekvenca besednih oblik v besedilih je namreč zelo različna. Zato je slovar črkovalnika, zgrajen na osnovi besedil (korpusa) običajno boljši od tistega, ki je zgrajen na osnovi nekega (knjižnega) slovarja.

Poleg izbire besed je zelo pomembna tudi njihova pravilnost. To sicer lahko dosežemo na več načinov, vendar je najzanesljivejše kar ročno preverjanje besed. Ker je tako preverjanje zelo počasno, smo si izdelali poseben vnosni program, ki to delo vseeno nekoliko pohitri.
Za kvaliteto črkovalnika je torej zelo pomemben postopek vnašanja novih besed, ki smo ga razdelili v tri osnovne faze:

- Zbiranje novih besed, večinoma iz besedil (korpus), iz katerih dobimo abecedno in frekvenčno urejen spisek "neznanih" besednih oblik oz. besed. Ta spisek trenutno vsebuje več kot 300.000 osnovnih besed oz. preko 1.000.000 besednih oblik, možnih kandidatov za vnos v slovar črkovalnika;
- Izbor najpogostejših besed iz spiska neznanih besed, ki jih še ni v slovarju;
- Preverjanje besed preko Pravopisa in Slovarja slovenskega knjižnega jezika ter vnos preko posebej prirejenega programa v vseh njenih pojavnih (pregibnih/morfoloških) oblikah v slovar.

S tako zgrajenim slovarjem smo do sedaj izdelali kar precejšnje število črkovalnikov v treh pojavnih oblikah - samostojen črkovalnik (uBesAna za DOS in Windows), dodatni modul (slovenski WordStar 7 za DOS, WordPerfect 6 in 7 za Windows, slovenski Word 6 za Windows) in vgrajena verzija (slovenski Word 7, 8 in 9 za Windows, Lotus Notes 4.x). V razvoju so še verzije za Corel WordPerfect 8 in Quark Express.

3 DELILNIK

Delilnik je drugi jezikovni modul, ki je običajno sestavni del različnih urejevalnikov. Glavni problem deljenja (zlogovanja) slovenskih besed so pravzaprav ohlapna pravila v pravopisu, ki način deljenja v večini primerov prepuščajo "akustičnemu občutku" posameznika. Ker je ta od človeka do človeka različen, je več tudi različnih deljenj določene besede, ki pa vsa ustrezajo obstoječim pravilom. Vsekakor med Slovenci prevladuje mnenje, da je "pravilno" deljenje potrebno in da deljenje na poljubno izbranem mestu ni v redu.
Na osnovi raziskave, v kateri je sodelovalo večje število "jezikovno usposobljenih" posameznikov, smo izdelali algoritem, ki poleg pravopisno predpisanih pravil upošteva tudi fonetične kriterije deljenja. Algoritem ima to lastnost, da poleg "najboljšega" deljenja predlaga tudi "možna" deljenja, ki so prav tako pravilna.
Tak primer je npr. beseda »nastaviti«, ki jo lahko delimo kot »na-sta-vi-ti« ali pa »nas-ta-vi-ti«. Amebisov delilnik predlaga obe možnosti (»na-s-ta-vi-ti«), pri čemer predlaga tudi boljšo (»na--s-ta--vi--ti«).
Seveda ima delilnik tudi slabosti in napake, ki jih na osnovi odkritih napačnih deljenj sproti popravljamo, bodisi v samem algoritmu, bodisi v dodatnem slovarju, ki ga lahko program uporablja.

Ker sta deljenje in zlogovanje zelo tesno povezana, smo algoritmu dodali še opcijo zlogovanja. Tako lahko isto funkcijo poleg uporabe v modulu za deljenje uporabimo tudi v drugih programih, npr. pri sintezi govora, kjer je od števila zlogov odvisna tudi hitrost izgovorjave določene besede.
Delilnik deluje kot vgrajeni modul v slovenskem Microsoft Word 7, 8 in 9 za Windows, v pripravi oz. razvoju pa je še dodatni modul za Corel WordPerfect.

4 BESEDNA ANALIZA

Besedna analiza je namenjena morfološki razčlembi besed v besedilu. Izvaja se s pomočjo slovarja in posebnih funkcij, katerih glavna naloga je odkrivanje nekaterih slovničnih napak v besedilu.
Programski del vsebuje naslednje glavne module:

- **Lematizator** - podfunkcija morfološkega analizatorja, ki poišče iz poljubne besedne oblike njeno osnovno besedo;
  pišemo --> pisati
  drevesoma --> drevo
  lepi --> lep, lepiti

- **Morfološki analizator** - modul, ki analizira poljubno besedno obliko in vrne njene osnovne morfološke informacije;

to
- prid. zaim. spol:srednji št.:ednina skl.:tožilnik dol:0 vrsta:41
- prid. zaim. spol:srednji št.:ednina skl.:imenovalnik dol:0 vrsta:41
- prid. zaim. spol:ženski št.:ednina skl.:orodnik dol:0 vrsta:41
- prid. zaim. spol:ženski št.:ednina skl.:tožilnik dol:0 vrsta:41

je
- glagol biti število:ednina oseba:3
- os. zaimek oseba:3 št.:ednina skl.:rodilnik spol:ženski obl.:1
- glagol jesti število:ednina oseba:3 <8>

cilj
- sam. spol:moški število:ednina sklon:tožilnik
- sam. spol:moški število:ednina sklon:imenovalnik

- **Generator pregibnih besednih oblik** (sklanjanje, spreganje) - modul, ki za določeno osnovno besedo generira vse njene besedne oblike;

Jaka
- Vrsta imena: osebno ime
- Spol: moški
- Svojilni pridevnik: Jakov
- Sklanjatev: 1052

Jaka
Jake, Jaka
Jaki, Jaku
Jako, Jaka
Jaki, Jaku
Jako, Jakom

Jaki, Jaka
Jak, Jakov
Jakama, Jakoma
Jaki, Jaka
Jakah, Jakih
Jakama, Jakoma

Jake, Jaki
Jak, Jakov
Jakam, Jakom
Jake, Jake
Jakah, Jakih
Jakami, Jaki

- **Skladenjski analizator** - modul, ki na osnovi določenih pravil ugotavlja skladnost besed in s tem določene napake v stavku.

To je življenski[1] cilj Novak Janeza[2].
1 /NAPAČNA BESEDA/ Nasvet: življenjski
2 /IZPADLA BESEDA/ <NAPAČEN VRSTNI RED IME : PRIIMEK>

S[1] njim grem v kamnik[2] in v[3] Krvavec.
1 /IZPADLA BESEDA/ <NAPAČNA OBLIKA PREDLOGA>
2 /IZPADLA BESEDA/ <MALA ZAČETNICA>
3 /IZPADLA BESEDA/ <NAPAČEN PREDLOG PRI LASTNEM IMENU>

To se je zgodili[1] na velikemu[2] vrtu[3].
1 /IZPADLA BESEDA/ <NEUJEMANJE GL. BITI : OP. DEL.>
2 /IZPADLA BESEDA/ <NEUJEMANJE S PREDLOGOM>
3 /IZPADLA BESEDA/ <NEUJEMANJE PRID. : SAM.>

Ni obiskal oddelke[1] XVX[2], XVV[3] iin[4] XIV.
1 /IZPADLA BESEDA/ <ZANIKANJE S TOŽILNIKOM>
2 /NEZNANA BESEDA/
3 /NEZNANA BESEDA/
4 /NEZNANA BESEDA/

Dal ga je Micki Novaku[1].
1 /IZPADLA BESEDA/ <NEUJEMANJE IME:PRIIMEK>

Po vrsti popravite vse kar želite[1].
1 /IZPADLA BESEDA/ <V STAVKU VERJETNO MANJKA VEJICA>

Slovarski del vsebuje preko 42.000 morfološko opisanih osnovnih besed, iz katerih lahko ustvarimo preko 1.500.000 različnih pomenskih oblik. Vsebina opisa neke besede je odvisna od besedne vrste. Slovar razlikuje med dvanajstimi besednimi vrstami: samostalniki, vezniki, medmeti, glagoli, členki, krajšave, pridevniki, predlogi, števniki, lastna imena, prislovi in zaimki.
Tako kot pri črkovalniku, je tudi tukaj pomembno število besed, ki jih slovar vsebuje, njihov izbor in pravilnost. Tudi tukaj je zato zelo pomemben postopek gradnje slovarja, ki ima naslednje faze:

- Zbiranje novih besed, večinoma iz besedil (korpus), iz katerih dobimo abecedno in frekvenčno urejen spisek "neznanih" besed. Ta spisek trenutno vsebuje več kot 300.000 osnovnih besed, možnih kandidatov za vnos v slovar;
- Izbor najpogostejših besed iz spiska neznanih besed, ki jih še ni v slovarju;
- Preverjanje besed preko Pravopisa in Slovarja slovenskega knjižnega jezika ter vnos preko posebej prirejenega programa v slovar.

Vse funkcije, ki jih besedna analiza omogoča, smo združili v programu BesAna [1], ki pa je še vedno dostopen le v DOS okolju. Razvoj Windows verzije se žal zaradi različnih razlogov še ni končal, smo pa že precej daleč.
Posamične module besedne analize s pridom uporabljamo tudi pri razvoju drugih programov s področja jezikovnih tehnologij, kot sta npr. program za strojno (podprto) prevajanje in sinteza govora.
Možnosti besedne analize so omejene, zato se lotevamo tudi prave stavčne analize, ki poleg morfoloških podatkov pri besedah v slovarju zahtevajo še dodatne sintaktične informacije. Ker je osnova za polnjenje slovarja s temi informacijami zadosti velika količina pravilnih stavkov (povedi), je prvi pogoj za uspešen razvoj stavčne analize tudi ustrezen korpus, zato smo se tega dela lotili najprej.

5 TEZAVER

Tezaver (tezavrus) je slovar z osnovnimi besedami in njihovimi medsebojnimi pomenskimi povezavami. Poleg sopomenk (sinonimov), združenih v pomenskih skupinah, in protipomenk (antonimov), vsebuje še povezave na referenčne besede.
Ker je podjetje Microsoft v svoj urejevalnik Word 8 želelo vgraditi slovenski tezaver, ki pa ga Slovenci žal nimamo niti v knjižni obliki, smo se lotili tudi tega oreha. V zelo kratkem času smo morali z zelo omejenimi sredstvi narediti nekaj, kar bi bilo vsaj v grobem podobno pravemu tezavru. Kot osnovo smo vzeli slovar sinonimov, ki je vgrajen v programu BesAna. V dveh (!) mesecih smo poleg programskega dela izdelali osnovno verzijo tezavra z okrog 75.000 med seboj povezanimi besedami. Seveda bo potrebnega še veliko dela s čiščenjem obstoječe baze in dodajanjem novih besed, ki jih trenutno ni v slovarju.
Primer zapisa nekaterih gesel:

agresija	0 napad, napadalnost, nasilje
	4 obramba
agresiven	3 bojevit, napadalen, nasilen, vsiljiv
	4 pasiven
	5 agresivnost
Legenda:	0 - samostalnik
	1 - razno
	2 - glagol
	3 - pridevnik
	4 - protipomenka
	5 - referenčna beseda

Ker obstoječa struktura tezavra v Wordu ni primerljiva s »pravimi« tezavri v drugih jezikih, smo se lotili razvoja novega tezavra na povsem novih temeljih. Kot osnovo smo vzeli standard ISO-2788. Tezaver po tej specifikaciji npr. uporabljajo programi za iskanje po besedilnih bazah podjetja Oracle. Za gradnjo tezavra smo razvili poseben program, ki omogoča dodajanje besed, kontrolo nad že dodanimi povezavami, dodajanje novih povezav ter brisanje obstoječih. Dodali smo tudi relacijo, ki je standard ISO-2788 ne vsebuje, v slovenščini pa je nekako potrebna. To je povezava med spoloma. Tako lahko npr. dodatno opišemo povezavo med besedama »natakar« in »natakarica«, ki sicer ne bi bila možna. Razvoj novega tezavra je še bolj na začetku. Predvidevamo, da bo ob sedanji hitrosti gradnje končan v nekaj letih. Ta čas bi seveda lahko tudi ustrezno skrajšali, če bi bilo v njegov razvoj vključeno večje število ljudi, kar pa je odvisno tudi od interesa zunanjih partnerjev.
Primer zapisa nekaterih gesel:

glasbilo
BT zvočilo
NT brenkalo
NT elektronsko glasbilo
NT godalo
NT harfa
NT klavirsko glasbilo
NT orgle
NT pihalo
NT tolkalo (glasbilo)
NT tradicionalno glasbilo
NT trobilo (glasbilo)

godalo
BT glasbilo
NT kontrabas
NT viola
NT violina
NT violončelo

sesalec (žival)
BTG vretenčar
NTG stokovec
NTG višji sesalec
NTG vrečar

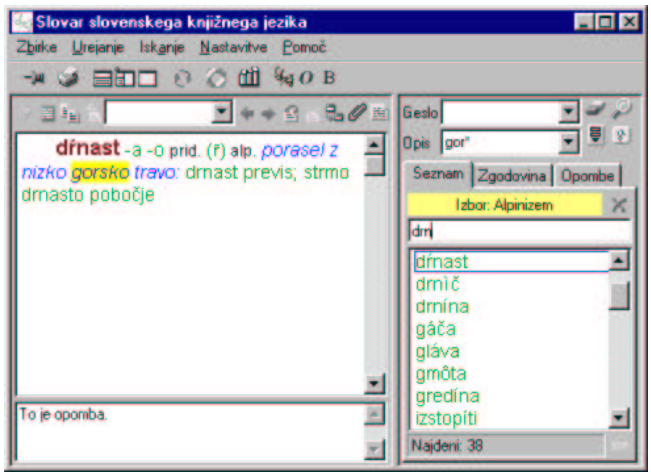
minuta
BT časovna enota
BTP ura (časovna enota)
NTP sekunda

6 KORPUS

Korpus je zbirka besedil, ki so osnova za različne jezikovne obdelave. Za potrebe lastnega razvoja besedne in stavčne analize smo pred leti razvili sistem ABIS, ki je skrbel za vnos in hranjenje besedil ter njihovo avtomatsko obdelavo. Tako zbrani korpus je vseboval preko 15 milijonov obdelanih besed. Pri vsem tem pa ni bil največji problem tehnična izvedba, ampak predvsem zagotavljanje zadostne količine materiala, ki je moral biti urejen tudi s pravnega vidika (urejeno varstvo avtorskih pravic). Zbrana besedila so nam služila predvsem pri nadaljnjem razvoju besedne analize, še posebej pa pri razvoju stavčne analize, ki je nadgradnja besedne analize in ki že upošteva nekatere sintaktične vidike celotnega stavka oz. povedi.
Medtem so tudi drugi, ki se ukvarjajo z jezikovnimi tehnologijami spoznali, kako koristen je lahko ustrezno zgrajen korpus besedil, zato smo v letu 1997 skupaj z DZS d. d., IJS in Filozofsko fakulteto pričeli s projektom FIDA [2], katerega cilj je zgraditi referenčni korpus slovenskega jezika. S skupnimi močmi gradimo doslej največjo zbirko slovenskih besedil, pri čemer sodeluje tudi veliko število drugih založb in založniških hiš, ustanov in podjetij ter drugih, ki prispevajo svoja (objavljena) besedila.
Korpus FIDA je zgrajen tako, da upošteva mednarodna priporočila za zapis besedil (standard SGML [3, 4], priporočila TEI [5]), programski del pa temelji na okolju Windows, ki je pri nas trenutno najbolj razširjen operacijski sistem. Za zdaj je to predvsem interni projekt štirih partnerjev, ki bo verjetno kasneje dostopen tudi širšemu krogu zainteresiranih uporabnikov, predvsem preko Interneta.
Sodelovali smo tudi pri gradnji večjezikovnega korpusa v okviru Copernicus projekta MULTEXT-EAST [6], ki nam bo služil predvsem pri razvoju sistema za strojno (podprto) prevajanje.

7 ELEKTRONSKI SLOVARJI

Trenutno najaktualnejše področje našega delovanja so elektronski slovarji. Skupaj z založbo DZS d.d., Založništvo literature, smo izdelali 5 elektronskih slovarjev v formatu ASP, ki so v prejšnjih letih izšli na disketah. Program ASP, s katerim lahko uporabljamo vse slovarje, je bil razvit za okolji DOS in Windows.
Ker so se v tem času precej uveljavili 32-bitni sistemi Windows, prav tako pa se je CD-ROM medij dodobra uveljavil, nadaljujemo z razvojem 32-bitne različice programa ASP, ki spremlja nove verzije slovarjev, dostopne na ploščah CD-ROM. Prvi tak slovar je Slovar slovenskega knjižnega jezika z Odzadnjim slovarjem slovenskega jezika in Besediščem slovenskega jezika z oblikoslovnimi podatki (SAZU in ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, DZS d. d., Založništvo literature, Amebis d. o. o.), v pripravi pa je izdelava vseh starih 16-bitnih slovarjev v novi 32-bitni različici, prav tako pa tudi izdelava novih slovarjev, ki jih do sedaj še ni bilo. Tudi novi sistem ASP32 je razvit tako, da lahko z enim programom uporabljamo več slovarjev, pri čemer je omogočena hkratnost uporabe, ter večja povezljivost med različnimi slovarji na eni strani, ter programom ASP32 in drugimi programi (npr. urejevalniki besedil) na drugi. Dodane so tudi nekatere nove funkcije, ki predvsem jezikoslovcem omogočajo precej naprednejšo uporabo slovarjev (konkordance, zahtevno iskanje, iskanje po frazah...).



8 STROJNO (PODPRTO) PREVAJANJE

Z vključevanjem Slovenije v Evropsko skupnost je tudi vedno večja potreba po pravem »prevajalniku«, ki naj bi omogočal še hitrejše delo oz. pomoč pri prevajanju besedil v tuje jezike in obratno. Ker po eni strani obvladujemo tehnologijo klasičnih računalniških slovarjev, po drugi strani pa mehanizme besedne oz. stavčne analize, smo se lotili tudi gradnje sistema za strojno (podprto) prevajanje. Ker je razvoj takega sistema izredno zahteven, smo se problema lotili v več stopnjah.

Prva verzija prevajalnika Presis prevaja iz angleščine v slovenščino. Prevajalno jedro programa je ločeno od uporabniškega vmesnika. V načrtu sta vmesnika za Internet in Word, ki pa ju za zdaj puščamo ob strani, saj je še precej dela s samim jedrom programa. To v prvi fazi temelji na enostavni menjavi besed izvornega besedila z ustreznimi besedami iz vgrajenega splošnega slovarja, pri čemer se upoštevajo le osnovne slovnične posebnosti izvornega in ciljnega jezika. Upoštevajo se tudi v slovar vgrajene fraze, kar izboljša kvaliteto prevoda. Tak prevod je sicer še vedno bolj podoben »indijanščini«, vendar nekomu, ki izvornega jezika niti malo ne razume, posreduje vsaj toliko informacij, da lahko oceni, o čem besedilo pravzaprav govori in ali ga je potrebno pravilno prevesti. Že tak sistem, ki ga ima večina ostalih evropskih jezikov, je lahko koristno orodje, ki nam prihrani veliko časa in truda.

Naslednji koraki za nadgradnjo programa Presis bodo v vedno večji meri upoštevali slovnično zgradbo obeh jezikov. Hkrati bomo dodajali nove pare izvornega in ciljnega jezika (slo/ang, nem/slo, slo/nem) ter po možnosti dodatne terminološke slovarje oz. filtre.

Seveda se je že takoj na začetku postavila dilema, ali razvijati lasten sistem, ali pa navezati stike z ustrezno tujo programsko hišo, ki že izdeluje podobne programe, ter skupaj z njo razviti slovenski modul. Obe možnosti imata svoje dobre in slabe strani. Odločili smo se za razvoj lastnega programa, pri čemer bomo pridobili potrebno specifično znanje, ki nam bo kasneje lahko koristilo pri morebitnem sodelovanju z drugimi.

9 SINTEZA GOVORA

Tudi zvočna obdelava jezika je eno izmed področij, ki smo se ga lotili. Ker se je v času naše odločitve z razpoznavo govora ukvarjalo že več skupin, smo se odločili za sintezo govora, ki je bila odrinjena na stranski tir. V pol leta smo izdelali sistem, ki smo ga tudi nekajkrat predstavili, vendar nas je prevelika količina razvojnega dela (tudi na drugih področjih) in pritisk po preživetju prisilil, da smo se za nekaj časa odrekli tako intenzivnemu razvoju in se posvetili povsem komercialnim potrebam trga. Tako se je razvoj na tem področju skoraj za dve leti povsem ustavil. Šele v zadnjem času spet povečujemo število ur, namenjeno razvoju sinteze govora. Stari sistem, ki je deloval še v okolju DOS smo prenesli v okolje Windows, pri čemer smo zagotovili delovanje v realnem času z osmimi glasovi hkrati, pri tem pa ohranili vse ostale prednosti starega sistema (nastavljivost višine, hitrosti in barve govora, dodatni izhodni filtri).

10 ZAKLJUČEK

Glede na stanje izpred desetih let se je v tem času marsikaj spremenilo. Kljub temu, da zaradi objektivnih razlogov nismo uspeli narediti vsega, kar smo želeli, pa je marsikaj že narejenega. Poleg osnovnih jezikovnih modulov, ki so postali že naša vsakdanjost, se je precej premaknilo tudi na bolj zahtevnih področjih.

Posebej očiten je napredek na področju sodelovanja z ostalimi skupinami v Sloveniji, ki se ukvarjajo z jezikovnimi tehnologijami. Še pred nekaj leti so nam ob želji po sodelovanju vsi, razen redkih izjem, obračali hrbet. Danes sodelujemo že v kar nekaj domačih in tujih projektih, ki združujejo različne razvojne skupine.

Še vedno pa za nas ostaja odprt problem, ki nam onemogoča intenzivnejše raziskovalno in razvojno delo, ki ga druge raziskovalne skupine na fakultetah in inštitutih nimajo. Kot komercialno podjetje namreč nimamo realnih možnosti pridobiti dodatna sredstva za domače raziskovalne projekte. Takega problema pri tujih projektih nimamo (npr. Copernicus MULTEXT-EAST). Tako moramo početi še vrsto drugih stvari, s katerimi lahko financiramo razvoj na področju jezikovnih tehnologij.

11 VIRI

[1] Šimunović, M., Holozan, P., Grilc, I., Romih, M. BesAna: Navodila za uporabo. Amebis. Ljubljana. 1993.

[2] Krek, S., Stabej, M., Gorjanc, V., Erjavec, T., Romih, M., Holozan, P. FIDA: korpus slovenskega jezika. http://fida.net

[3] Batagelj, V. (1995). Uvod v SGML. http://vlado.mat.uni-lj.si/vlado/sgml/sgmluvod.htm

[4] Goldfarb, C.F. (1990). The SGML Handbook. Clarendon Press, Oxford, 1990.
[5] Ide, N., Veronis, J. (ur.). (1995). The Text Encoding Initiative: Background and Context. Kluwer Academic Publishers, Dordrecht.
[6] Erjavec, T., Ide, N. The MULTEXT-East Corpus. V Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98. Granada. str. 971-974, 1998.