Infrastruktura za razvoj jezikovnih tehnologij - korpus FIDA in sistem ASES

Miro Romih, Peter Holozan

Amebis d.o.o.
Bakovnik 3, 1241 Kamnik, Slovenija
miro.romih@amebis.si, peter.holozan@amebis.si

Za razvoj jezikovnih tehnologij je ključnega pomena ustrezna infrastruktura. To lahko sestavljajo bolj ali manj urejene oz. popolne zbirke besed, besedil, izgovorjav, ter različni slovarji, leksikoni, tezavri itd. Brez zadostnega števila ustrezno urejenih jezikovnih podatkov si razvoja in izdelave splošno uporabnih programov ter sistemov s področja jezikovnih tehnologij res ne moremo predstavljati.

Osnovno infrastrukturo (z izjemo govornih tehnologij) predstavljajo besedila oz. besedilni korpusi. Iz samih besedil v elektronski obliki, če ta niso enotno urejena in zadosti obdelana, prav veliko podatkov ne moremo pridobiti. Lahko jih npr. uporabimo kot osnovo za širjenje baze črkovalnika, izluščimo nekaj statističnih podatkov, to pa je tudi vse. Za bolj temeljito analizo in s tem dostop do večjega števila podatkov o jeziku pa je potrebno zgraditi korpus.

FIDA je uravnotežen referenčni korpus slovenskih besedil, ki je nastal v sodelovanju štirih partnerjev. S svojimi 100 milijoni besed je v tem trenutku največji slovenski korpus, z ustreznimi orodji za iskanje in dodatne obdelave konkordančnih spiskov pa zagotovo najmočnejše orodje za analizo slovenskega jezika. Poleg same zasnove in organizacije zbiranja ustreznih besedil je bilo težišče dela na pretvarjanju besedil v enoten format zapisa, oblikoslovnem označevanju teh besedil, izdelavi iskalnega programa in orodij za dodatne obdelave, ter izdelavi vmesnika za dostop do korpusa.

Seveda pa do vseh jezikovnih podatkov, ki so potrebni za izdelavo zahtevnejših programov s področja jezikovnih tehnologij, ni moč priti le s pomočjo korpusa. Z njim si lahko pomagamo pri razvoju sintaktičnega analizatorja, do ostalih podatkov, ki jih npr. potrebuje prevajalni program, pa moramo priti na drugačen način in v ta namen zgraditi drugačno jezikovno bazo. In če jo že gradimo, je seveda najbolje, če je baza ena sama, v njej pa so zbrani vsi podatki, ki jih različni programi potrebujejo.

Aktivni slovenski elektronski slovar (ASES) je podatkovna zbirka, ki vsebuje različne podatke o slovenskem jeziku in tudi drugih tujih jezikih. Osnovne enote sistema ASES so med seboj povezani pojmi, preko katerih se slovenske besede povezujejo z besedami v drugih jezikih. Poleg teh povezav se pojmi med seboj lahko povezujejo tudi v različne druge skupine ter v enega ali več delnih tezavrov. Pojmi poleg nekaterih pomenskih in drugih statističnih informacij vsebujejo še povezave na ustrezne besede oz. besedne zveze, sinonimne in antonimne povezave itd. Same besede vsebujejo osnovne morfološke informacije ter podatke o zlogovanju in izgovorjavi.

ASES torej združuje pojmovnik, enega ali več tezavrov, besednjak, slovar sinonimov, enojezični, več dvojezičnih in večjezični splošni slovar, terminološke slovarje, slovar zlogovanja, leksikon izgovorjav in še kaj. Večino podatkov je v ASES ob sicer veliki pomoči ustreznih podzbirk in programov še vedno potrebno vpisati ali vsaj preveriti ročno, zato je gradnja tako velikega sistema izredno počasna in zahteva veliko jezikovnega znanja. Ker pa se jezik nenehno spreminja, je naloga še toliko težja. Vendar ko enkrat vnesemo zadostno količino podatkov, lahko z ustreznimi orodji iz take baze enostavno in hitro izločimo podatke, ki jih v danem trenutku potrebujemo za določen program, npr. črkovalnik, delilnik, sintetizator govora ali prevajalni sistem.