Jezikovne tehnologije

Hiter razvoj računalništva in informatike je posegel na številna področja človekovega udejstvovanja. Tudi obdelava naravnih jezikov se temu ni mogla izogniti. Tako smo dobili številna orodja, brez katerih si danes težko zamišljamo tudi povsem običajna opravila, kot je npr. pisanje.

Miro Romih

Preprosto povedano so jezikovne tehnologije vse tiste informacijske tehnologije, ki se ukvarjajo z naravnimi jeziki, med katere sodi tudi slovenščina. Jezikovne tehnologije postajajo vse pomembnejši sestavni del informacijskih tehnologij, to pa je v osnovi omogočila šele dovolj velika procesna moč sodobnih računalnikov in drugih naprav, ki so hkrati sposobne hraniti velikansko število informacij. Naravni jeziki so namreč izjemno velika zakladnica besed, zapletenih pravil in znanja, ki ga je človeštvo ustvarjalo tisoče let. Zaradi zahtevnosti bi zato jezikovne tehnologije zlahka uvrstili ob bok drugim, bolj znanim visokim tehnologijam, kot so npr. nanotehnologija, telekomunikacije, umetna inteligenca in robotika.

Naštejmo nekaj značilnih primerov, kako lahko jezikovne tehnologije vplivajo na naše vsakdanje življenje. Črkovalnik ali celo slovnični pregledovalnik sta že danes marsikomu nepogrešljiv pripomoček, brez katerega si je težko zamišljati pisanje besedil. Prevajanje besedil iz tujih jezikov in obratno v veliki meri poteka ob pomoči katerega od elektronskih slovarjev, programa za pomoč pri prevajanju ali kar neposredno s pomočjo strojnega prevajanja. V zadnjem času je precej govora o pametni hiši prihodnosti, kjer je ena izmed pomembnejših funkcionalnosti tudi govorna komunikacija z napravami. Marsikje so vlogo človeškega operaterja prevzeli govorni telefonski odzivniki. Pri iskanju koristnih informacij si že danes lahko pomagamo s katerim od programov, sposobnih komuniciranja v naravnem jeziku; nekateri izmed teh sistemov postajajo celo sestavni del najnovejših modelov avtomobilov. Slepi in slabovidni lažje uporabljajo marsikatero napravo ali računalniški program, ki je posebej zanje prilagojen z dodatnim govornim vmesnikom (ŽIT 2002/7–8, str. 44). Vsega tega in še marsičesa drugega brez razvoja jezikovnih tehnologij ne bi bilo. Na eni strani nam torej jezikovne tehnologije pomagajo pri naših običajnih opravilih v zvezi z jezikom (npr. pravilno pisanje, hitrejše in lažje prevajanje), po drugi strani pa nam lahko približajo uporabo nekaterih naprav, še zlasti določenih funkcij računalnika, na nam prijaznejši način (npr. uporaba govora).

Na področju jezikovnih tehnologij deluje veliko število podjetij, organizacij in ustanov, razvijalci in proizvajalci se srečujejo na mnogih mednarodnih konferencah, kakovost številnih rešitev ter sistemov s tega področja je vsaj za večino glavnih svetovnih jezikov na precej visoki ravni. Slovenija oz. slovenščina sicer nekoliko zaostaja za najrazvitejšimi, vendar je hkrati v prednosti pred številnimi narodi oz. jeziki, ki jo po številu »uporabnikov« krepko presegajo. Razvojne skupine v Sloveniji so združene v Slovenskem društvu za jezikovne tehnologije (SDJT), ki vsaki dve leti organizira posebno konferenco na to temo. Člani društva sodelujejo v številnih mednarodnih organizacijah in projektih, tako da je zagotovljen stalni stik z razvojnimi dosežki v tujini.

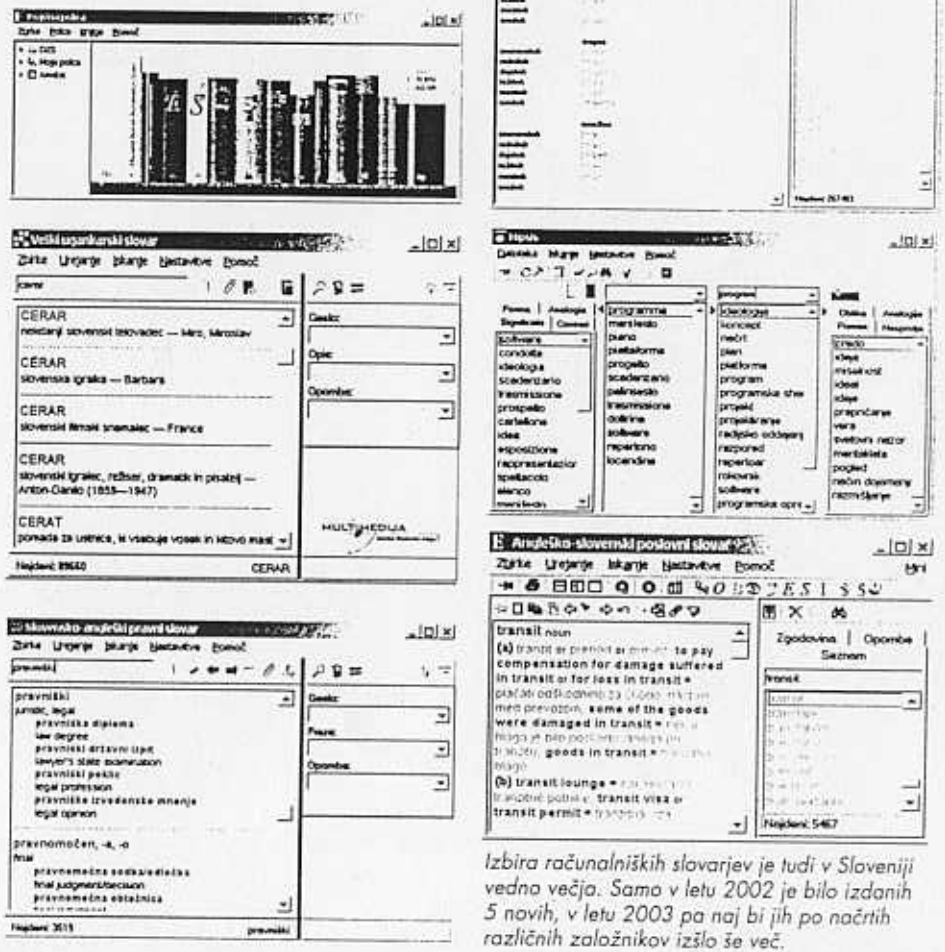
STRUKTURA JEZIKOVNIH TEHNOLOGIJ

Jezikovne tehnologije obsegajo vrsto področij in značilnih (računalniških) aplikacij, njihovo zgradbo, medsebojno povezanost in

Temelj jezikovnih tehnologij pomenijo različni korpusi in aktivni elektronski slovarji.

odvisnost pa si lahko ponazorimo na različne načine. Vsekakor je poleg razvoja informacijske tehnologije osnovni razlog za obstoj jezikovnih tehnologij obstoj naravnih jezikov samih, sistema pisave ter množica pisanih in govorjenih besedil, dostopnih v raznih oblikah, predvsem elektronski.

Nadgradnja tej neurejeni množici besedil in hkrati temelj jezikovnih tehnologij pomenijo t. i. korpusi, ki prinašajo urejenost ter s tem možnost nadaljnjih avtomatiziranih obdelav. Ker pa samo urejena besedila ne morejo biti sama po sebi osnova za



Izbira računalniških slovarjev je tudi v Sloveniji vedno večja. Samo v letu 2002 je bilo izdanih 5 novih, v letu 2003 pa naj bi jih po načrtih različnih založnikov izšlo še več.

delo različnih jezikovnih orodij, je treba korpus nadgraditi s t. i. aktivnimi elektronskimi slovarji, ki vsebujejo dodatne jezikovne informacije, pridobljene bodisi na osnovi analize korpusa, bodisi z ročnim vnosom na osnovi pridobljenega jezikovnega znanja. Z zadostno količino podatkov in znanja o jeziku, vsebovanem v aktivnih elektronskih slovarjih, lahko ti postanejo osnova za različna jezikovna področja in njihove aplikacije.

V nadaljevanju si nekoliko podrobneje oglejmo nekatera bolj znana področja jezikovnih tehnologij.

ELEKTRONSKI SLOVAR

S pasivnimi elektronskimi slovarji oz. računalniškimi slovarji se je verjetno srečal že vsakdo izmed nas. Mogoče bi bolj sodili na področje elektronskega založništva, pa vendar jih zaradi vsebinske narave in prepletenosti z drugimi področji jezikovnih tehnologij prištevamo tudi k slednjim. Enojezične, dvojezične in večjezične računalniške slovarje, ki so v bistvu le elektronska izvedba klasičnih knjižnih slovarjev, tudi pri nas že dolgo uporabljamo kot pomoč pri pisanju in prevajanju besedil. Z njihovo pomočjo je delo lažje in predvsem hitrejše. Še zlasti so nepogrešljivi pri uporabi več slovarjev hkrati ter iskanju informacij po vsebini.

Število dostopnih domačih slovarjev v elektronski obliki se iz leta v leto povečuje, njihova kakovost pa izboljšuje. Razvoj slovarskih orodij gre predvsem v smeri enostavnejše in hitrejše uporabe ter integracije slovarjev z drugimi orodji, kot so npr. prevajalniki. Močan razvoj pa poteka tudi v smeri gradnje t. i. aktivnih elektronskih slovarjev. Ti niso prvenstveno namenjeni ljudem, temveč predvsem programom in aplikacijam, zato je pri njih osnovna zahteva, da so strojno berljivi. Tak slovar je npr. sestavni del prevajalnega sistema in je v osnovi zgrajen drugače kot običajni računalniški slovar.

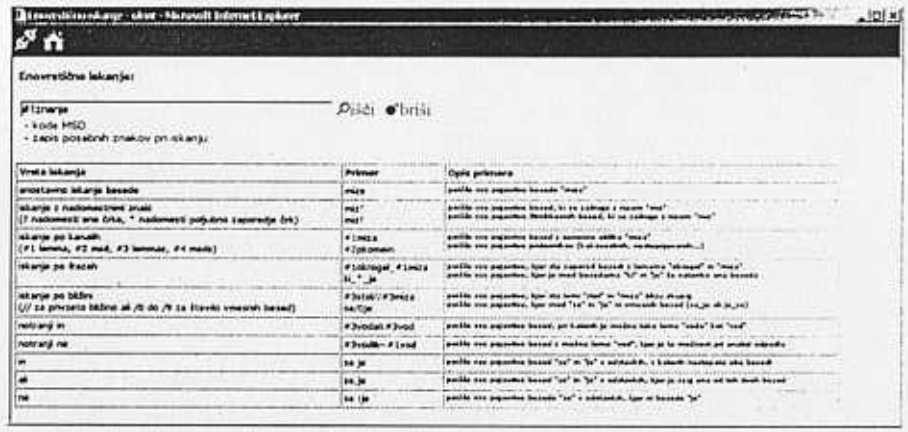
KORPUS

Ustrezna infrastruktura je ključnega pomena za razvoj jezikovnih tehnologij. To lahko sestavljajo bolj ali manj urejene oz. popolne zbirke besed, besedil in izgovarjav 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 zamišljati. Za gradnjo aktivnih (strojno berljivih) elektronskih slovarjev, ki vsebujejo različne informacije (znanje) o jeziku, so bistvenega pomena besedilni korpusi. To so zbirke besedil, urejene po določenih pravilih in pretvorjene v enotno obliko zapisa, ter ustreznih programov, s katerimi preiskujemo besedila in nad njimi izvajamo določene statistične operacije. V zadnjih letih se je težišče jezikoslovnega raziskovalnega dela zelo prevesilo v smeri besedilnih korpusov, saj so ti – še zlasti, če so zelo obsežni in pravilno uravnoteženi –, edini pravi pokazatelj dejanskega stanja jezika. Poleg besedilnih obstajajo tudi korpusi govorjenih besedil, ki se uporabljajo kot infrastruktura za razvoj razpoznave in sinteze govora. Ti so manjši od večine besedilnih korpusov, saj je na voljo manj posnetih zvočnih materialov kot samih besedil, njihova obdelava pa je precej zahtevnejša in dolgotrajnejša.

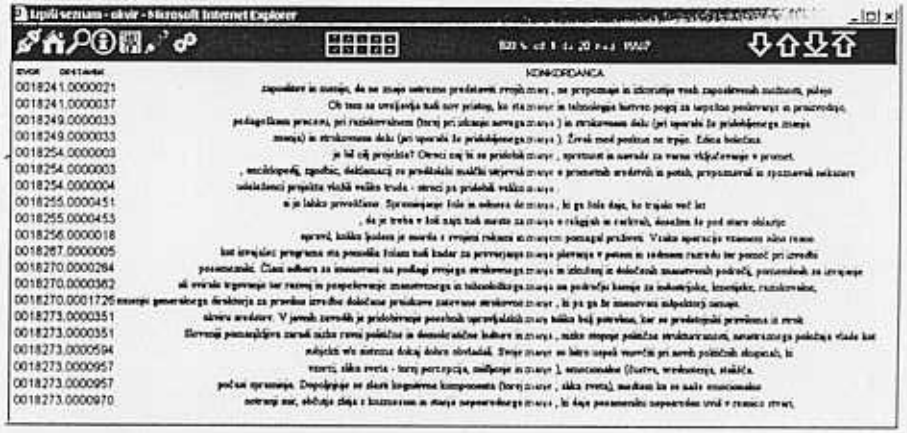
Sodobni besedilni korpusi so zelo veliki in navadno obsegajo nekaj sto milijonov besed. Gradnja tako obsežnih korpusov je seveda zahtevna in dolgotrajna. Poleg same organizacije pridobivanja ustreznih besedil, ki pa je neprimerno manjši problem kot pred leti, je treba rešiti tudi vprašanje avtorskih pravic. Dobljena besedila je nato treba iz različnih vhodnih formatov pretvoriti v en sam končni format zapisa ter ob tem v glavo besedila zapisati osnovne podatke o izvoru. Ker zaradi različnega namena uporabe obstaja več različnih vrst korpusov (zgodovinski, govorni, referenčni, totalni, ...), je treba temu primerno prilagoditi tudi zbiranje materialov in način gradnje. Za potrebe referenčnega korpusa, ki naj bi predstavljal čim bolj verodostojno sliko stanja jezika, je treba zbrane materiale na podlagi določenih pravil uravnotežiti. Za uravnoteženje se uporabljajo določena razmerja glede lektorske obdelave (lektorirano, nelektorirano), izvora (dnevniki, tedniki, revije, knjige, ...) in zvrsti besedil (drame, leposlovje, proza, ...) tako, da ti predstavljajo pravo sliko jezika. Če korpus ni referenčni,

Uporaba korpusa

Besedilni korpus lahko uporabimo v različne namene, med drugim kot koristno orodje pri razvoju slovarjev. Kot primer je prikazan postopek iskanja besed v korpusu FIDA, ki se najpogosteje neposredno vežejo s samostalnikom *znanje*. Te podatke lahko uporabimo pri izdelavi slovarskega gesla kot primere najpogosteje uporabljanih zvez.



1. korak: Iskanje vseh pojavljanj besede *znanje* v korpusu. Ker lahko ta nastopa tudi v različnih sklonih, moramo vpisati ustrezen iskalni pogoj, da iskalni program vse oblike tudi upošteva.



2. korak: Ko dobimo vsa pojavljanja besede *znanje*, uporabimo funkcijo iskanja prve besede na levi.

3. korak: Dobljene rezultate uredimo po ustrezni statistični funkciji. Rezultat je po pogostosti urejen seznam besed, ki nastopajo v slovenščini neposredno pred besedo *znanje*.

uravnoteženje ni potrebno. Naslednji korak pri gradnji korpusa je običajno še oblikoslovno označevanje besed, nazadnje pa je treba korpus spraviti v obliko, ki je primerna za hitro iskanje in različne statistične operacije nad njim. Zaradi velikosti korpusov so ti v glavnem nameščeni na strežnikih, uporabniki pa do njih dostopajo prek spleta.

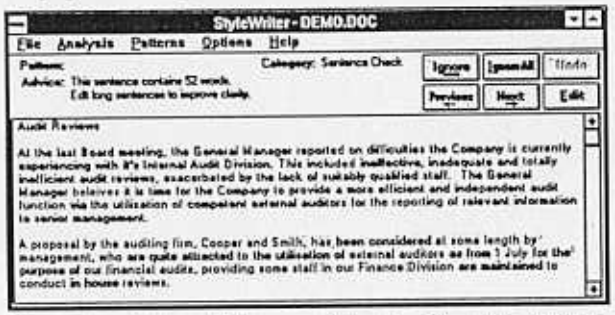
Večina jezikov že ima zgrajene različne vrste korpusov, ki pa se seveda nenehno dograjujejo. Tudi slovenski jezik ima vrsto različnih korpusov. Ta čas največji uravnotežen referenčni besedilni korpus slovenskega jezika je korpus FIDA, ki vsebuje dobrih 100 milijonov besed, kar je zelo dobra osnova za različne jezikovne raziskave. V sodelovanju številnih domačih partnerjev pa se pripravlja tudi nov projekt za izgradnjo še večjega jezikovnega korpusa, ki bo omogočil še večji napredek na področju jezikoslovja in jezikovnih tehnologij.

ČRKOVALNIK

Črkovalnik je najosnovnejši jezikovni modul in je običajno sestavni del urejevalnikov besedil. Njegova naloga je odkrivanje tipkarskih napak v besedilu, njegovo delovanje pa je v osnovi zelo preprosto. S pomočjo slovarja pravilnih besednih oblik preveri, ali določena beseda obstaja. Če besede v slovarju ne najde, jo proglasi za napačno. Seveda je lahko beseda povsem pravilno zapisana, vendar slovar te oblike ne vsebuje. Zato je pri črkovalniku zelo pomembno zadostno število pravilnih oblik, ki jih vsebuje slovar. Navadno je to nekaj sto tisoč besednih oblik, pri pregibnih jezikih, med katere sodi tudi slovenščina, pa se to število lahko povzpne tudi krepko čez milijon. Ker v njem seveda nikdar ni vseh besed, saj te neprestano nastajajo in se spreminjajo, črkovalnik običajno omogoča tudi dodajanje novih besednih oblik v dodatni slovar. Poleg tega ima lahko še nekaj drugih funkcij, s katerimi pomaga uporabniku pri delu, kot so npr. nasveti pri napačnih besedah. Poleg urejevalnikov besedil črkovalnik pri svojem delu s pridom uporabljajo tudi drugi programi, kot so npr. programi za optično prepoznavanje znakov (OCR), ki z njegovo pomočjo skušajo samodejno izločiti čim več napak pri razpoznavanju znakov, zato so lahko tudi precej bolj zanesljivi.

SLOVNIČNI PREGLEDOVALNIK

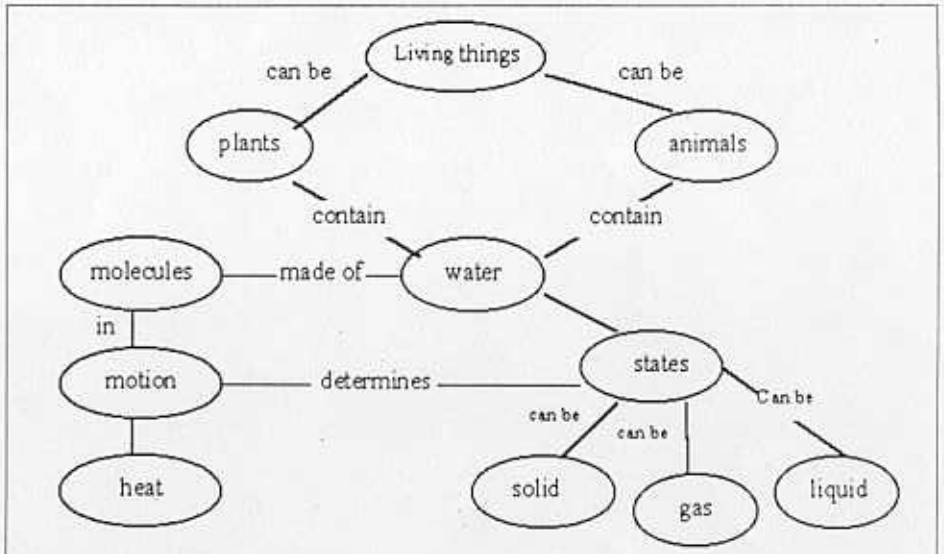
Slovnični pregledovalnik je naslednji jezikovni modul, ki pa v urejevalnikih besedil ni tako pogosto zastopan. V nasprotju s črkovalnikom, ki preverja le očitne tipkarske napake, slovnični pregledovalnik preverja sintakso celega stavka, zato odkrije tudi slovnične napake, ki jih črkovalnik ne zna. Seveda je izdelava tega modula neprimerno zahtevnejša, predvsem pa jezikovno zelo odvisna. Že samo izdelava ustreznega slovarja je zahtevno in dolgotrajno delo. Vsaka beseda mora namreč vsebovati številne informacije, ki so odvisne od njene besedne vrste. Pri samostalniku npr. potrebujemo informacijo o spolu, vse pripadajoče oblike za različna števila in sklone, svojilni pridevnik (če obstaja), podatek o živosti – in še kaj bi se našlo. Poleg slovarja je treba za skoraj vsak jezik posebej zgraditi tudi sintaktični analizator, ki uporablja prej omenjeni slovar in ki je jedro tega modula. Običajno je naloga slovničnega pregledovalnika tudi svetovanje pri odpravi napake.



Slovnični pregledovalniki so sposobni poleg tipkarskih odkrivati tudi zapletenejše slovnične napake.

TEZAVER

Tretji jezikovni modul, ki je tudi pogosto del urejevalnikov, je tezaver. Ta posebni slovar



V semantični mreži so posamezni pomeni med seboj povezani z različnimi vrstami odnosov, ki odražajo naravo sveta, v katerem živimo.

med seboj povezuje različne besede oz. pomene, ki jih te besede predstavljajo. Koristi nam v primerih, ko pri pisanju potrebujemo soroden izraz, sinonim, protipomenko ali referenčno besedo (npr. pridevnik samostalnika, žensko obliko, nadpomen). Tezaver pomeni neke vrste semantično mrežo, v kateri povezave med pomeni kot osnovnimi elementi tezavra odsevajo naravo sveta, v katerem živimo. Informacije, ki so zbrane v tezavru, lahko pri svojem delu koristno uporabljajo tudi nekateri drugi programi s področja jezikovnih tehnologij. Kot primer lahko vzamemo sistem za strojno prevajanje, ki v primeru, da za kak pomen nima prevoda, lahko uporabi prevod njegovega nadpomena, če ta seveda obstaja.

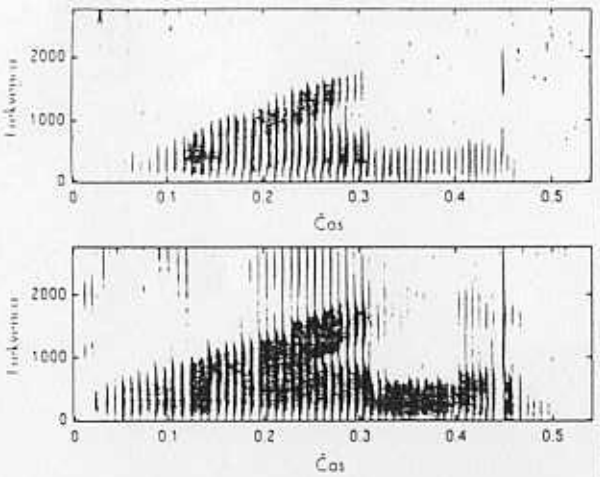
DELILNIK

Delilnik za večino jezikov sodi med preprostejše jezikovne module. V urejevalnikih v primeru, ko je treba čim bolj zapolniti vrstico besedila, poskrbi, da se ustrezna beseda deli na pravem mestu. Običajno delilnik pri svojem delu uporablja poseben jezikovno odvisen algoritem, pri izjemah pa si pomaga z dodatnim slovarjem, ki ga uporabniki lahko tudi sami dopolnjujejo. Delilnik oz. modul za zlogovanje poleg urejevalnikov koristno uporabljajo še drugi programi. Npr. sintetizator govora ga s pridom uporablja pri tvorjenju govora, saj je hitrost izgovarjave med drugim odvisna tudi od števila zlogov v besedi.

RAZPOZNAVA GOVORA

Govorne tehnologije so zelo poznano področje jezikovnih tehnologij. V grobem jih razdelimo na razpoznavo in sintezo govora. Vsaka ima svoje razvojne probleme in možnosti, obema pa je cilj vzpostaviti govorno komunikacijo med ljudmi in različnimi napravami – seveda le tam, kjer je to smiselno in uporabno. S pomočjo razvitih modulov že delujejo nekatere splošno znane rešitve, kot so spletni agenti in informacijski portali.

Pretvorba govora v besedilo je bila na začetkih računalništva le tiha želja razvijalcev, danes pa je že povsem običajna funkcija marsikaterega sistema ali naprave. Nekateri boljši sistemi precej dobro razpoznavajo celo neprekinjeni govor. Glavni razlog za hiter razvoj tega področja v zadnjih letih je v zadostni procesorski in pomnilniški moči sodobnih naprav, kar je omogočilo splošno uporabnost razpoznavalnikov govora na povsem običajnih računalnikih in s tem njihovo popularizacijo.



Osnova za predstavitev zvočnega zapisa besede v računalniku so spektrogrami, ki dajejo časovno in frekvenčno sliko govornega signala. Zgornji spektrogram kaže zapis izgovorjene angleške besede word, spodnji pa računalniško generirano enako besedo.

Postopek razpoznave govora v osnovi temelji na primerjanju kratkih delov govora z vzorci, ki so v posebni podatkovni bazi. Seveda se pred tem po posebnem postopku obdelajo (pretvorijo iz časovnega v frekvenčni prostor), potem pa se s pomočjo različnih modelov izvršita primerjava in sestavljanje manjših enot v večje. Končni cilj je seveda pretvorba govornega signala v besedilo s čim manj napakami. Pri tem so odločilni nekateri dejavniki, ki pomembno vplivajo na kakovost razpoznave. Med njimi je npr. spremljajoči šum ali zvok okolice, ki ga človek na nam neznani način uspešno »izloča«, medtem ko imajo umetni sistemi z njim precej težav. Podobno je z menjavo govorca, ki se mu človek v trenutku prilagodi, pa čeprav ga sliši prvič, sistemi za razpoznavo govora pa pri tem postanejo precej manj zanesljivi. Kljub številnim težavam je razvoj omenjenih sistemov v precejšnjem razmahu, zato lahko na tem področju še veliko pričakujemo.

SINTEZA GOVORA

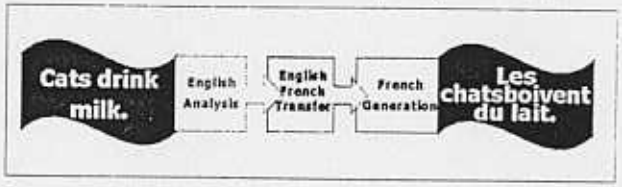
Sinteza govora je jezikovno precej bolj odvisna od izbranega jezika kot razpoznava. Zato je prvi pogoj za uspešno pretvorbo besedila v govor izgradnja ustrezne baze izgovarjav in algoritmov. Pod sintezo govora uvrščamo predvsem tiste rešitve, kjer ni vnaprej posnetih celih besed ali celo stavkov, ampak se govor generira sproti, z lepljenjem osnovnih glasov. Nekatere rešitve, ki ta čas prevladujejo, govor tvorijo z lepljenjem sestavljenih glasov, ker so rezultati glede na vložek najbolj optimalni. Seveda obstajajo tudi druge rešitve, ki npr. simulirajo govorni trakt in dajejo zelo dobre končne rezultate. Žal je časovno krmiljenje takih sistemov izjemno zahtevno, razvoj pa izredno dolgotrajen ter drag. Vsekakor se sinteza govora tako kot razpoznava iz laboratorijev počasi vse bolj seli k uporabnikom, z njeno pomočjo pa številne naprave postajajo vedno bolj prijazne za uporabo, kar je glavni cilj jezikovnih tehnologij.

STROJNO (PODPRTO) PREVAJANJE

Strojno (podprto) prevajanje postaja eno izmed pomembnejših področij jezikovnih tehnologij, zato mu na tem mestu posvetimo malo več pozornosti. V času vsesplošne globalizacije je za ohranjanje določenega jezika obstoj takih sistemov posebej pomemben. To še zlasti velja za tako majhen jezik, kot je slovenščina. Velika večina prevajalnih sistemov je namenjena predvsem podpori prevajanju, kar pomeni, da so le v pomoč prevajalcu pri njegovem delu, še vedno pa je prevajalec tisti, ki besedilo dokončno prevede. Seveda ob izredno hitrem napredku strojne in programske opreme nekateri prevajalni sistemi postajajo tako dobri, še posebej na ozko specializiranih področjih, da lahko govorimo že kar o izključno strojnem prevajanju, saj poseg prevajalca pogosto sploh ni več potreben.

Pod izrazom »strojno podprto prevajanje« razumemo prevajanje, ki ga v prvi vrsti izvaja človek prevajalec ob podpori različnih računalniških orodij, med katerimi igrajo pomembno vlogo t. i. pomnilniki prevodov. Ti so jezikovno neodvisni, delujejo pa tako, da prevajalcu poskušajo pomagati na podlagi prevodov celih stavkov, ki jih je za kolikor toliko uspešno prevajanje treba prej vnesti v prevajalno bazo. In teh nikakor ni malo. Zato se za splošno prevajanje taka vrsta programa zdi precej neuporabna, se pa toliko bolj uporabljajo na omejenih področjih, kjer se določeni stavki precej ponavljajo (npr. navodila za uporabo, zakonodaja). Pomnilniki prevodov imajo vgrajene posebne mehanizme približnega iskanja, brez katerih bi bili v večini primerov povsem neuporabni. Tako pa lahko prevajalcu tudi v primeru, da konkretnega stavka ni v prevajalni bazi, predlagajo stavek in njegov prevod, ki je po obliki oz. uporabljenih besedah najbližji. Prevajalec mora seveda prevod ustrezno spremeniti, kar pa je običajno še vedno precej hitreje, kot če bi moral stavek prevesti v celoti.

Strojno prevajanje poteka brez pomoči človeka prevajalca. Taki sistemi so jezikovno odvisni, njihov razvoj pa zahteva izredno veliko znanja in časa. Brez poprejšnjega učenja so sposobni prevesti poljubno besedilo, kar je velika prednost pri njihovi uporabi. V splošnem njihov prevod ni vedno najboljši, zato so pregled in popravki s strani uporabnika še vedno zaželeni.



Strojni prevajalnik najprej analizira stavek v izvornem jeziku, v drugem koraku prevede posamezne dele stavka, na koncu pa oblikuje stavek po pravilih ciljnega jezika.

Prevajalne sisteme s področja strojnega prevajanja bi lahko razvrstili po različnih merilih. Če kot merilo vzamemo »globino« oz. način dela, na kakršnega vsebinsko prevajajo, bi jih lahko v grobem razdelili v štiri skupine:

1. Prevajalni sistemi, ki le menjajo besedno obliko za besedno obliko. Taki sistemi so za pregibne jezike, kakršen je tudi slovenščina, v splošnem neuporabni. V poštev bi lahko prišli le pri zelo omejenem področju prevajanja.

2. Prevajalni sistemi, ki znajo vhodno besedilo lematizirati (poiskati osnovno obliko besede). Tak sistem brez kakršne koli dodatne logike bi npr. za slovenščino že dal kolikor toliko dobre rezultate. Seveda pa v splošnem stavki ne bi bili posebno berljivi, ker prevajanje še vedno poteka le na podlagi zamenjave besede za besedo, pa tudi sintaksa izvornega in ciljnega jezika se ne upošteva.

3. Prevajalni sistemi, ki poleg morfologije upoštevajo tudi sintakso obeh jezikov. Prevod takega sistema je v splošnem že kar lepo berljiv, ker pa sistem nima pomenske analize, prevod vsebinsko ne ustreza vedno najbolj.

4. Prevajalne sisteme, ki prevajajo na podlagi morfološke, sintaktične in pomenske analize. Ti dajejo pri prevajanju najboljše rezultate, a seveda niso nezmotljivi. Da bi pravilno »razumeli« stavek, ki ga prevajajo, zaradi boljše analize del informacij prenašajo iz prejšnjih stavkov, saj se pogosto zgodi, da se kak stavek nanaša na prejšnjega ali prejšnje stavke. Osnovna enota prevajanja so tako odstavki. Tudi rezultati prevajanja so običajno boljši pri daljših stavkih, saj tam prevajalni sistem lahko natančneje določi pomen posameznih besed kot pri krajših stavkih.

V splošnem se prevajanje izvaja v treh osnovnih korakih.

Najprej se z analizo poskuša čim bolj natančno ugotoviti, kakšen je pomen posamezne besede v izvornem stavku ter kako so te med seboj pomensko povezane. Včasih – še zlasti pri na videz preprostih stavkih – se zdi, da analiza ni posebej zahtevna stvar. Vendar ni tako. Analizator mora najti in upoštevati vse mogoče pomenske in sintaktične možnosti, saj nikoli ne moremo natančno vedeti, katera od njih je prava: ali tista, ki je sicer najobičajnejša, ali katera druga. Kot primer vzemimo stavek *To je Miro Cerar*. Na prvi pogled ni nobenega dvoma. Pomen stavka si lahko razlagamo, kot da je to oseba moškega

spola z imenom in priimkom Miro Cerar. Vendar pa si isti stavek lahko razlagamo še v dveh drugih pomenih, ki sicer nista običajna, sintaktično in pomensko pa sta povsem možna. Stavek lahko razumemo tudi tako, kot da to (nekaj) jé žensko z imenom in priimkom Mira Cerar, ali pa, kot da oseba z imenom Miro Cerar jé to (nekaj). Če bi imeli le sintaktični analizator, potem so vse tri pomenske možnosti sintaktično pravilne in enakovredne. Šele z dodatno pomensko analizo, ki si pomaga z določenimi statističnimi informacijami, pridobljenimi s pomočjo analize posameznega jezika (korpusa), ter informacijami, pridobljenimi na podlagi prejšnjih stavkov, lahko te možnosti utežimo po verjetnosti ter s tem določimo najverjetnejši prevod.

Ko tako izluščimo določen pomen posameznih besed ali fraz v izvornem stavku, moramo osnovne oblike teh besed ustrezno pretvoriti v besede in fraze ciljnega jezika. Pri tem si prevajalnik običajno pomaga z vgrajenim splošnim slovarjem. Ker ima lahko določena beseda ali fraza več možnih prevodov, je treba vse prenesti naprej do tretjega modula, ki se potem odloči za »pravi« prevod. Poleg osnovnega splošnega slovarja ima prevajalnik lahko še dodatne terminološke slovarje, osebne slovarje uporabnika ali vzorce prevodov celih stavkov, katerih naloga je čim bolje in s čim več možnimi besedami prevesti v ciljni jezik glede na področje, s katerega je prevajano besedilo. Vsako področje ima namreč svoje zakonitosti tvorjenja stavkov in specifične prevode določenih besed.

Ko je prevajalnik opremljen s podatki o pomenih posameznih besed in njihovih možnih prevodih, mora opraviti še tretjo fazo prevajanja, tj. sintezo. Ta ima nalogo, da oblikuje stavek po slovničnih pravilih ciljnega jezika, ki so lahko bistveno drugačna od pravil izvornega jezika. Poleg tega se mora odločiti za najboljši prevod ustreznega pojma. Pri tem lahko preprosto vrne tisti prevod, ki je v splošnem (na podlagi statistične analize) najpogostejši, lahko pa na podlagi drugih besed v stavku (na podlagi statistične analize kolokacij) določi prevod, ki je v primeru konkretnega stavka verjetno boljši. Vsekakor ima statistična analiza obeh jezikov, tako izvornega kot ciljnega, bistveno vlogo pri kakovostnem prevajanju. To je v bistvu tisto znanje, ki se pri človeku (prevajalcu) odraža v obliki izkušenj, in na podlagi katerih se odloča tudi sam. Niso torej dovolj le veliki in dobri slovarji prevodov ter algoritmi za stavčno analizo in sintezo; prevajalnik mora vsebovati čim bolj popolno sliko o obeh jezikih, da se lahko kar najbolj približa dobremu prevodu.

Prevajalni sistemi

Od tujih prevajalnih sistemov so pri nas najbolj poznani in uporabljani izdelki podjetij Trados in Systran, ki pa so za slovenščino praktično uporabni le kot pomnilniki prevodov. Najnovejši izdelek na področju strojnega prevajanja slovenščine je prevajalni sistem *Presis* podjetja Amebis iz Kamnika. V osnovi je to »pravi« prevajalnik, ki pa ima vgrajen pomnilnik prevodov, tako da je uporaben za različne vrste uporabnikov. Za zdaj je na voljo le slovensko-angleški prevajalni modul, če bo šlo vse po načrtih, pa bo *Presis* že čez kako leto prevajal tudi v obrnjeni smeri, torej iz angleščine v slovenščino. Da bi čim bolj zadostil potrebam uporabnikov, *Presis* v odvisnosti od kupljenega paketa vsebuje več programov z različnimi uporabniškimi vmesniki, ki pa uporabljajo isto prevajalno jedro. Ker zasnova celega sistema oz. njegovih programov temelji na strukturi odjemalec–strežnik, se lahko prevajalno jedro po dogovoru integrira v različne aplikacije (spletne, mobilne, mrežne itd.).

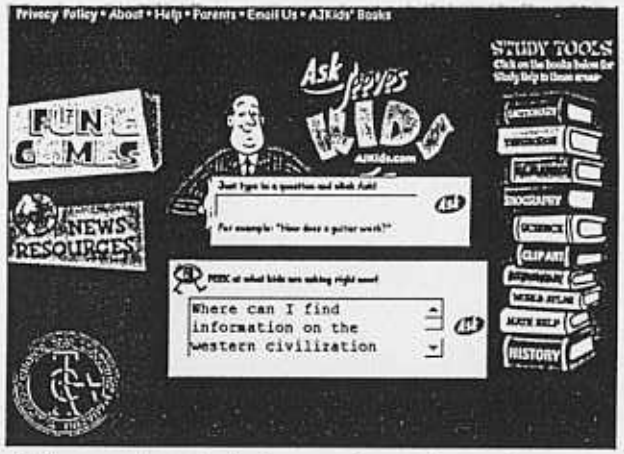
Presis PRO je najzmogljivejši program prevajalnega sistema Presis.

Najnovejši prevajalni sistemi poskušajo kombinirati prednosti tako »čistih« prevajalnikov kot pomnilnikov prevodov. Tako dajejo uporabniku možnost, da po želji uporablja tiste mehanizme, ki so zanj najugodnejši glede hitrosti in kakovosti prevajanja. Hitrosti prevajanja, ki jih zmorejo sodobni prevajalni sistemi, so zelo odvisne od strojne in programske opreme, na kateri tečejo, navadno pa so sposobni prevesti nekaj deset strani (formata A 4) besedila na minuto. To pa zlasti pri prevajanju večjih količin besedila seveda lahko prinese velikanske prihranke časa.

Težav, ki jih je treba rešiti pri tako zahtevnem opravilu, kot je prevajanje, je precej. Postopno izboljševanje različnih delov prevajalnega sistema, npr. slovarskega dela, ki že zaradi spreminjanja jezikov verjetno ne bo nikoli končano, pa bo prej ali slej prineslo rezultate, s katerimi bodo zadovoljni tudi najzahtevnejši uporabniki.

DIALOG V NARAVNEM JEZIKU

Naravna komunikacija človeka z drugimi ljudmi poteka v obliki govora, za pismene tudi prek pisnih sporočil. V prvem in drugem primeru sporazumevanje poteka v obliki stavkov, oblikovanih po pravilih določenega jezika. Ta način komunikacije nam omogoča posredovanje vsega, kar želimo nekomu sporočiti, enako sprejemati in se na osnovi tako sprejetih informacij učiti. Seveda pa za uspešnim dialogom stoji nekakšna »notranja logika«, ki jo ljudje brez vsakih težav obvladujemo, za računalnike pa pomeni zelo veliko oviro. Z njenim odstranjevanjem se ukvarja posebna veja jezikovnih tehnologij – razvoj sistemov dialoga. Napredek v tej smeri bi ljudem precej olajšal delo z računalniki oz. njim podobnimi napravami, saj bi bila z dodatno uporabo govornih tehnologij takšna komunikacija med človekom in računalnikom praktično enaka komunikaciji med ljudmi.



Ask Jeeves sodi med najbolj znane informacijske portale na spletu, ki komunicirajo v naravnem jeziku.

SKLEP

V tem prispevku našteta področja in rešitve seveda nikakor niso vse, kar bi lahko uvrstili k jezikovnim tehnologijam. Obstaja namreč še vrsta drugih, manj znanih področij, z razvojem katerih skušajo strokovnjaki uporabo sodobnih informacijskih naprav približati ljudem. S hitrim razvojem informacijske tehnologije pa se pred jezikovne tehnologije postavljajo tudi vedno novi in novi izzivi, na katere bo treba najti odgovor. Počakajmo in se pustimo presenetiti. Morda se večina tistega, o čemer danes le sanjamo, kmalu tudi uresniči.

http://...

nl.ijs.si/sdjt (Slovensko društvo za jezikovne tehnologije – SDJT)
www.fida.net (Slovenski referenčni korpus FIDA)
www.amebis.si (podjetje Amebis)
presis.amebis.si (prevajalni sistem Presis)
www.ask.com (informacijski portal Ask Jeeves)