K SAMODEJNEMU PRIDOBIVANJU JEZIKOVNIH VIROV S POMOČJO INTERNETA

Zdravko Kačič, Bogomir Horvat, Matej Rojc, Aleksandra Zögling Markuš*

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova 17, 2000 Maribor, Slovenija
{kacic,bogo.horvat}@uni-mb.si
*Univerza v Mariboru, Center za interdisciplinarne in multidisciplinarne raziskave in študije
Razlagova 22, 2000 Maribor, Slovenija

Povzetek

V članku predstavljamo zasnovo sistema za avtomatsko pridobivanje in obdelavo besedil. Sistem omogoča iskanje domačih strani na internetu in pridobivanje ustreznih besedil. Ta služijo kot osnova za gradnjo besedilnih korpusov, pri tem pa je moč ob morfološki analizi besedil dopolnjevati obstoječe pisne jezikovne vire, kot so: morfološki slovar, slovar lastnih imen in fonetični slovar. Proces gradnje jezikovnih virov je pri tem sestavljen iz dveh delov. V prvem sistem pridobi ustrezna besedila s pomočjo interneta, opravi razčlenitev besedil in besedno analizo ter tvori sezname nepoznanih besed. V drugem delu pa ekspert s pomočjo različnih programskih orodij analizira sezname in ustrezno dopolni besedilne korpuse ter druge obstoječe jezikovne vire. Definirani sistem zmanjšuje obseg potrebnega ekspertnega dela in tako pohitri in poceni gradnjo pisnih jezikovnih virov.

Abstract

In this paper, a system for automatic acquisition and analysis of text is described. The system enables search for appropriate www addresses on the Internet and acquisition of appropriate texts. On the basis of these, a text corpus is build. By morphologic analysis of the texts the existing language resources such as morphologic dictionary, pronunciation dictionary, and dictionary of proper names can be upgraded. The process of building language resources is here composed of two phases. In first, the system searches for proper texts over the Internet and performs basic extraction and annotation of the text corpus. By parsing the text the system generates a list of unknown items. In the second phase, the expert performs final annotation of the generated corpus and by analysing the lists of items he updates the existing resources. The defined system reduces the required expert workload and in this way speeds up the process of building the language resources.

1. Uvod

Razvoj sodobnih jezikovnih tehnologij zahteva pospešen razvoj ustreznih jezikovnih virov. Te v zadnjih letih še posebej intenzivno razvijajo v državah evropske unije, saj predstavljajo osnovno infrastrukturo za razvoj sistemov jezikovnih tehnologij, ki bodo omogočili preseganje jezikovnih pregrad v sodobni Evropi.

Jezikovne vire lahko delimo v govorne in pisne vire (H□ge,1998). Med glavne govorne jezikovne vire prištevamo: označene baze izgovarjav, fonetične slovarje in označene ter neoznačene besedilne korpuse. Označene baze izgovarjav in fonetični slovarji predstavljajo glavne govorne jezikovne vire pri razvoju sistemov avtomatskega razpoznavanja govora – za modeliranje na akustičnem nivoju. Besedilne korpuse uporabljamo predvsem za jezikovno modeliranje pri avtomatskem razpoznavanju govora in sistemih avtomatske sinteze govora. Pri slednjih uporabljamo pri izvajanju grafemsko-fonemske transkripcije tudi fonetične slovarje. Prav tako predstavljajo besedilni korpusi pomembne vire pri procesiranju naravnega jezika, kjer jih prištevamo med pisne jezikovne vire, kamor sodijo tudi različni leksikoni in drugi slovarji.

2. Razvoj jezikovnih virov za slovenski jezik

V slovenskem prostoru se je obsežnejši razvoj jezikovnih virov za potrebe področja jezikovnih tehnologij pričel v devetdesetih letih. Med raziskovalnimi in univerzitetnimi ustanovami se v Ljubljani z izgradnjo jezikovnih virov ukvarjajo predvsem na Fakulteti za elektrotehniko (bazi izgovarjav Gopolis, Luz) (Gros, 2000)(Dobrišek, 1998), Inštitutu Jožef Stefan (projekt Multext, korpus FIDA) (Erjavec, 1998)(Erjavec, 1998a), Filozofski fakulteti in SAZU (SSKJ). V Mariboru na Fakulteti za elektrotehniko, računalništvo in informatiko (Center za jezikovne tehnologije), Pedagoški fakulteti in Centru za interdisciplinarne in multidisciplinarne raziskave in študije Univerze v Mariboru (CIMRŠ). Med podjetji pa predvsem podjetje Amebis iz Kranja (Romih, 1998) v sodelovanju z DZS in akademskimi inštitucijami.

V okviru aktivnosti Centra za jezikovne tehnologije, ki od leta 1996 deluje na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru (FERI), je bilo v zadnjih petih letih v sodelovanju z Laboratorijem za digitalno procesiranje signalov in Laboratorijem za digitalne in informacijske sisiteme, oba s FERI, ter CIMRŠ, zgrajenih več jezikovnih virov.

V letih 1994-1998 je bila zgrajena baza izgovarjav SNABI, ki vsebuje govorni signal telefonske kvalitete 82 govorcev in govorni signal studijske kvalitete 56 govorcev. Pri tem je vsak govorec izgovoril v poprečju 200 stavkov, 80 izoliranih besed, številske nize in abecedo, 21 govorcev je prebralo 400 stavkov, 20 pa še odlomek iz izbranega teksta, s katerim so simulirali nalogo narekovanja (Kačič, 2000).

V letih 1995-1996 je bil v okviru projekta Copernicus Onomastica zgrajen fonetični slovar slovenskih lastnih imen. Slovar vsebuje imena in priimke, imena ulic, podjetij, ustanov in krajev. Zajema vsa lastna imena, ki so bila v letu 1995 zajeta v centralnem registru imen Statističnega urada Republike

Slovenije. Vsako enoto slovarja sestavljata ortografska transkripcija imena, do pet različic fonetične transkripcije imena, etimološki podatki in dodatni komentar. Slovar obsega nekaj več kot 282.000 enot (Kačič. 1998).

V letih 1997 in 1998 je bila v okviru projekta SpeechDat II posneta baza izgovarjav telefonske kvalitete 1000 govorcev (Kaiser, 1998). Baza predstavlja prvo bazo za slovenski jezik, ki omogoča razvoj sistemov avtomatskega razpoznavanja govora za delo v realnem okolju. Baza je uravnotežena glede na vrsto kriterijev, kot so: spol, starost, narečno področje, izobrazba in snemalno okolje. Z njeno pomočjo sta bila razvita sistema govornega dialoga O-tel in InfoDesk. Sistem O-tel je večjezični sistem obratnega telefonskega imenika, ki je izveden za slovenski, nemški in angleški jezik (sistem je bil uspešno predstavljen na prvi olimpijadi sistemov govornega dialoga v okviru konference Eurospeech, ki je bila leta 1997 na Rodosu, Grčija) (Miksić, 1998). Informacijski sistem InfoDesk (posredovanje podatkov o rezultatih preizkusov znanj) pa že dve leti uspešno deluje na FERI. V okviru aktivnosti SpeechDat Task Force (projekt COST 249 – Continuous Speech Recognition over the telephone) sodelujejo raziskovalci s FERI pri raziskavah na področju večjezičnega razpoznavanja govora (trenutno potekajo raziskave za 5 jezikov), kjer kot osnova služijo baze SpeechDat II (Johansen, 2000), (Lindberg 2000).

Ob koncu leta 1999 je v sodelovanju s podjetjem Telekom Slovenija stekel projekt snemanja baze izgovarjav PoliDat. Cilj projekta je posneti bazo izgovarjav telefonske kvalitete za 1000 govorcev ob uporabi ISDN telefonskega omrežja in 1000 govorcev preko GSM telefonskega omrežja. Specifikacije baze so skladne s specifikacijami baz SpeechDat II za ISDN telefonijo, upoštevane pa so tudi specifične zahteve za bazo GSM telefonskega omrežja.

V aprilu leta 2000 so bile posnete baze izgovarjav emocionalnega govora za slovenski in angleški jezik. Korpus baze izgovarjav slovenskega jezika vključuje 190 izoliranih besed in stavkov. Ti so bili izgovorjeni v šestih različnih emocionalnih stanjih in dveh nevtralnih. Bazo sta posnela igralka in igralec, ki sta vseh šest emocij posnela dvakrat, s štirinajstdnevnim premorom med snemanjima. Za angleški jezik so bazo na enak način posneli trije govorci in ena govorka. Korpus baze vključuje 186 izoliranih besed in stavkov. Baze bodo služile predvsem za analizo prozodičnih značilnosti govora in razvoj postopkov za njihovo modeliranje.

V okviru aktivnosti zagotavljanja pisnih tekstovnih virov je bil izdelan besedilni korpus, ki trenutno vsebuje 2 milijona besed prečiščenega in prebranega besedila. Celotna velikost korpusa pa je trenutno 32 milijonov besed. Pri označevanju korpusa je uporabljan SGML standard. Besedilni korpus je bil dosedaj uporabljen predvsem pri razvoju jezikovnih modelov pri sistemih avtomatskega razpoznavanja slovenskega govora (Sepesy Maučec, 1998).

Cilj zastavljenih aktivnosti Centra za jezikovne tehnologije pri gradnji pisnih in govornih jezikovnih virov je predvsem zagotoviti potrebne vire pri razvoju sistemov govorne tehnologije – sistemov avtomatskega razpoznavanja in sinteze govora ter sistemov govornega dialoga. Zato je zasnova pisnih jezikovnih virov (na primer besedilnih korpusov) z nekaterih vidikov morda nekoliko ožja, kot pa to zahtevata na primer področji leksikologije in leksikografije. Tudi pri označevanju korpusov so nekoliko bolj upoštevane zahteve, ki izhajajo iz potreb razvoja sistemov govorne tehnologije.

Pri razvoju sistemov procesiranja naravnega jezika predstavljajo predvsem pisni jezikovni viri osnovo uspešnemu razvoju. Izgradnja besedilnih korpusov, različnih slovarjev in drugih virov pa je zelo zamudno in obsežno opravilo, zahteva pa tudi velika finančna vlaganja predvsem zaradi potrebe po obsežnem angažiranju strokovnjakov s posameznih jezikoslovnih področij. Delna avtomatizacija procesa izgradnje pisnih jezikovnih virov bi lahko bistveno pripomogla k njihovemu hitrejšemu razvoju.

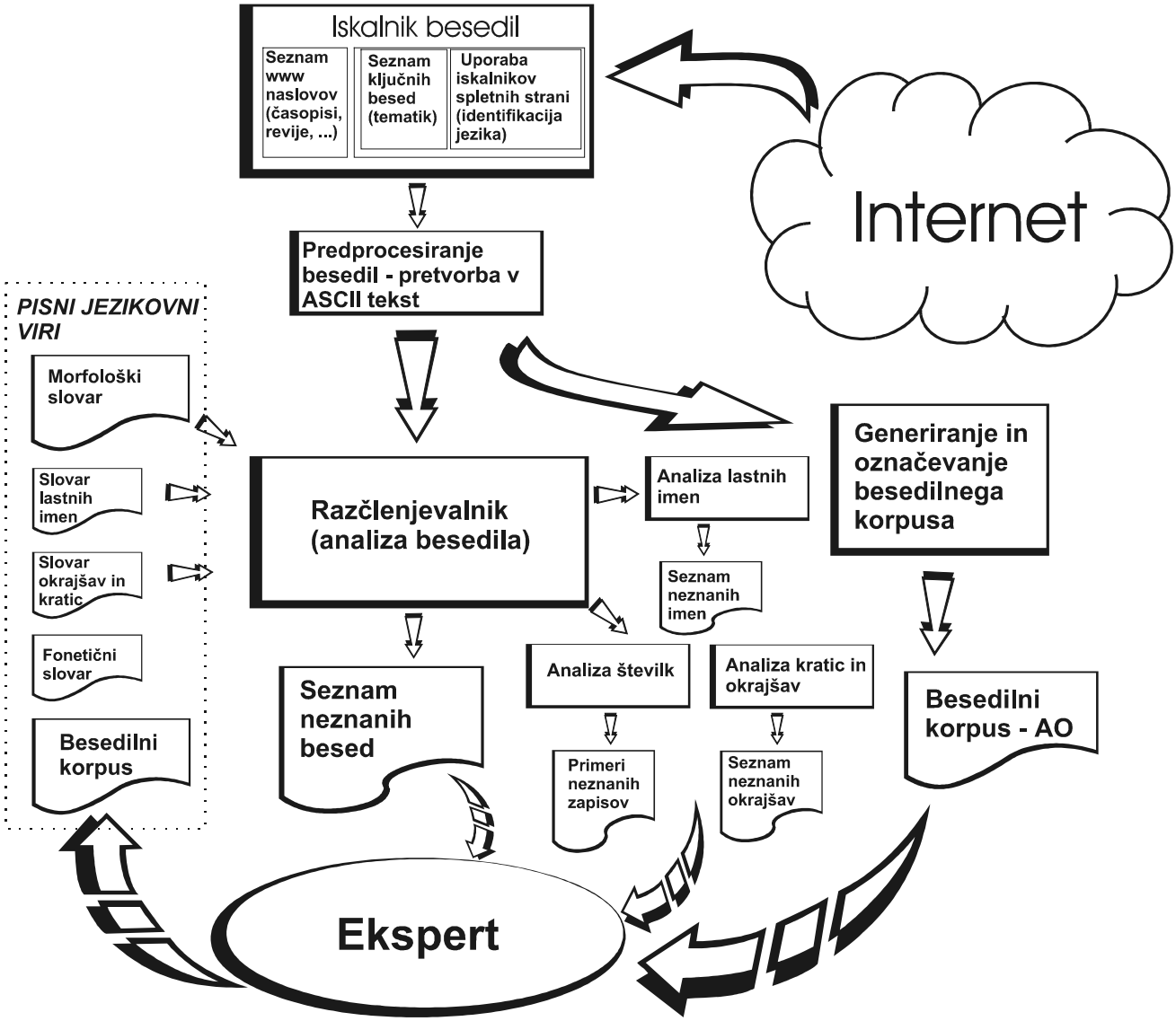
V preteklosti je bilo moč ustrezna besedila potrebna za gradnjo različnih pisnih jezikovnih virov večinoma dobiti le od založniških ali časopisnih hiš in je to predstavljalo eno od večjih ovir pri hitrem in uspešnem razvoju jezikovnih virov. Z razvojem interneta pa je dosegljivost različnih besedil postala mnogo enostavnejša. Obseg besedil nenehno narašča, kar daje internetu značaj globalnega informacijskega vira. Res, da založniške in časopisne hiše na internetu največkrat še ne objavljajo celotnih izdaj časopisov, revij ali knjig, vendar je obseg besedil že tolikšen, da ga je smiselno uporabiti pri gradnji različnih pisnih jezikovnih virov.

Uporaba interneta kot vira besedil pa omogoča izgradnjo sistema, s katerim lahko avtomatiziramo postopek pridobivanja besedil, njihovega prečiščevanja (predprocesiranja) in delno tudi analize ter samodejnega generiranja različnih virov. Na ta način lahko v precejšnji meri zmanjšamo potrebni obseg ekspertnega dela, saj je to nujno le v zadnji fazi - fazi preverjanja pravilnosti zapisa, ugotavljanju pravopisnih napak in obravnavi novih enot v različnih virih (npr. fonetični in morfološki slovar, slovar lastnih imen, slovar okrajšav in kratic).

V članku podajamo zasnovo sistema za avtomatsko pridobivanje in obdelavo besedil ter gradnjo pisnih jezikovnih virov. Sistem omogoča avtomatsko pridobivanje besedil z interneta, njihovo prečiščevanje, izgradnjo različnih pisnih jezikovnih virov in pripravo le-teh za preverjanje, ki ga v zadnji fazi opravijo eksperti.

3. Sistem za avtomatsko pridobivanje in obdelavo besedil

Sistem za avtomatsko pridobivanje in obdelavo besedil sestavlja množica orodij, ki omogočajo iskanje ustreznih dokumentov na internetu, njihovo predprocesiranje, analizo ter tvorjenje različnih seznamov besed in korpusov. Slika 1 prikazuje blokovno shemo sistema. Poglavitna naloga sistema je omogočiti čim hitrejše kreiranje in preverjanje pisnih jezikovnih virov ob čim manjšem vložku potrebnega ekspertnega dela. Glavni moduli sistema so: iskalnik besedil, modul predprocesiranja besedil in pretvorbe besedil v ASCII tekst, razčlenjevalnik besedila in modul generiranja in označevanja korpusov besedil. Slednja dva modula generirata sezname enot, ki jih v nadaljevanju preveri ekspert s pomočjo ustreznih programskih orodij ter jih doda obstoječim pisnim jezikovnim virom.



Slika 1: Sistem za avtomatsko pridobivanje in obdelavo besedil

Sistem tako omogoča ob gradnji besedilnega korpusa tudi dopolnjevanje drugih pisnih jezikovnih virov, kot so: slovar lastnih imen, fonetični slovar, slovar okrajšav in kratic ter morfološki slovar.

3.1. Iskalnik besedil

Naloga iskalnika besedil je iskanje domačih strani na internetu, ki vsebujejo ustrezna besedila v slovenskem jeziku. Pri tem sta najpreprostejši dve možnosti. Pri prvi iskalnik besedil išče ustrezna besedila na domačih straneh različnih ustanov, ki periodično objavljajo nove dokumente. To so predvsem založniške in časopisne hiše. Tu je način prijavljanja na domače strani in povzemanje besedil prilagojen organizaciji domačih strani in direktorijski strukturi, ki je pri tem uporabljena. Struktura se tu običajno ne spreminja, kar olajša iskanje novih dokumentov in poenostavlja izvedbo modula. Vendar je v tem primeru pridobivanje dokumentov omejeno le na manjše število virov.

Druga možnost je definiranje seznama ključnih besed (tematik) in uporaba iskalnikov spletnih strani. Iskanje besedil je v tem primeru bolj zapleteno. Ob upoštevanju seznama ključnih besed išče sistem s pomočjo različnih iskalnikov dokumentov (Alta Vista, Lycos, Yahoo …) dokumente, ki vsebujejo definirane ključne besede. Po dobljenem seznamu dokumentov, ki ga generira iskalnik, mora sistem dokumente iz seznama poiskati in jih razčleniti. Če so primerni, jih posreduje modulu predprocesiranja teksta. Pri tem mora sistem s pomočjo identifikacije jezika zagotoviti, da je obravnavani dokument napisan v slovenskem jeziku. Hkrati mora ažurirati bazo www naslovov in ime dokumenta, da ne bo večkrat analiziral isti dokument. Omenjeno iskanje dokumentov po internetu je uporabno predvsem v primerih, ko želimo zgraditi tematsko orientirane jezikovne vire. Takšne zahteve so prisotne pri gradnji sistemov govorne tehnologije (sistemi govornega dialoga, narekovalni sistemi) za posamezne aplikacijske domene (pravo, medicina, ekonomija), ko moramo zagotoviti ustrezne jezikovne modele. Pri tem lahko uporabnik seznam ključnih besed (tematik) poljubno določi.

V trenutno delujoči izvedbi iskalnika besedila le-ta vsakodnevno ob vnaprej določenem času avtomatsko poišče dokumente na vnaprej določenih naslovih domačih strani. Iskanje besedil s pomočjo seznama ključnih besed je trenutno še v fazi razvoja.

3.2. Modul predprocesiranja besedil

Naloga modula predprocesiranja besedil je pretvorba vhodnih besedil v ASCII tekst. Pri tem mora iz vhodnega besedila izločiti vse kontrolne znake, slike, in druge informacije o formatiranju dokumenta. Obdelano besedilo mora vključiti v besedilni korpus in posredovati razčlenjevalniku besedila v nadaljnjo analizo. Pri vključitvi besedila v besedilni korpus mora izvesti začetno označitev dokumenta ob uporabi standarda SGML. Trenutno izvedeni modul podpira formate vhodnih besedil: html, pdf in ASCII.

3.3. Razčlenjevalnik

Naloga razčlenjevalnika je morfološka analiza besedila ob iskanju novih enot (besed), ki še ne obstajajo v razpoložljivih pisnih jezikovnih virih (Rojc, 1999). Pri analizi uporablja morfološki leksikon, slovar lastnih imen (trenutno slovar Onomastica) in slovar okrajšav in kratic. Ob primerjanju vsebin omenjenih virov generira sezname enot, ki jih v razpoložljivih virih še ni (morfološkem slovarju, slovarju lastnih imen in slovarju okrajšav in kratic). Prav tako generira tudi seznam zapisov številk, ki jih razčlenjevalnik še ne zmore razčleniti.

3.4. Naloge eksperta

Dosedaj opisani postopek poteka popolnoma samodejno. Z ozirom na vnaprej določen časovni razpored lahko sistem samodejno išče dokumente na posameznih www naslovih in sicer dnevno, tedensko, mesečno … , jih obdela in zapisuje v ustrezne sezname. Naloga eksperta je, da s pomočjo različnih programskih orodij pregleda generirane sezname besed in jih uvrstiti v obstoječe jezikovne vire. Tako z uporabo programskega orodja morf (Rojc, 2000) obdela generirani seznam neznanih besed in pri tem ugotavlja ali so besede v seznamu napačno zapisane (posledica tipkarske napake v obravnavanem viru) ali pa dejansko predstavljajo besedo, ki še ni vsebovana v morfološkem slovarju. V tem primeru oblikoslovno označi besede, to med drugim vključuje tudi označevanje naglasa, zlogovanje in fonetično transkripcijo, ter besedo zapiše v morfološki in fonetični slovar. Podobno postopa pri analizi seznama lastnih imen, le da tu delo zajema le preverjanje pravilnosti zapisa imena (izločitev možnosti pravopisnih napak), izvedbo fonetične transkripcije in zapis v obstoječi slovar lastnih imen in fonetični slovar. Pri analizi seznama kratic in okrajšav preveri pravilnost kratice oziroma okrajšave in jo skupaj s celotnim izpisom kratice oziroma okrajšave zapiše v slovar okrajšav in kratic. Hkrati tvori tudi fonetični zapis in ga zapiše v fonetični slovar. Pri analizi neznanih zapisov številk (na primer, različne oblike zapisov datumov, denarnih zneskov in drugih številskih vrednosti) najprej preveri smiselnost posameznega zapisa in nato zapise združi v skupine podobnih zapisov. Takšna njihova razvrstitev služi pozneje razvijalcu pri dopolnjevanju modula analize številskih zapisov v sklopu razčlenjevalnika.

4. Zaključek

Definirani sistem avtomatskega pridobivanja besedil omogoča učinkovito gradnjo besedilnih korpusov in s pomočjo analize izhodnih seznamov, ki jih tvori modul razčlenjevalnika, tudi nenehno dopolnjevanje različnih jezikovnih virov (predvsem morfološkega in fonetičnega slovarja).

V prihodnosti nameravamo izboljšati in dopolniti možnosti iskanja besedil po internetu in izboljšati orodja, ki so na voljo ekspertu pri analizi generiranih seznamov besed in dopolnjevanju obstoječih pisnih jezikovnih virov.

5. Viri

Dobrišek, S., Gros, J., Mihelič, F., Pavešič, N. (1998) Recording and Labelling of the GOPOLIS Slovenian Speech Database. Proceedings of the First International Conference on Language Resources and Evaluation, pp. 1089-1096. Granada, Spain.

Erjavec, T., Gorjanc, V., Stabej, M. (1998) Korpus FIDA Zbornik konference Jezikovne tehnologije za slovenski jezik, Ljubljana, pp. 124-127.

Erjavec, T. (1998) The MULTEXT-East Slovene Lexicon. Zbornik konference ERK`98, Portorož, Slovenija, pp. 189-192.

Gros, J., Mihelič, F., Dobrišek, S., Erjavec, T., Žganec M. (2000) Corpora of Slovene Spoken Language for Multi-lingual Applications. Zbornik Second International Conference on Language Recources and Evaluation. Athens, pp. 953-956.

Höge, H. (1998) Spoken language resources for Voice Driven Man Machine Interfaces. Proceedings of the First International Conference on Language Resources and Evaluation, pp. 209-216. Granada, Spain.

Johansen, F.T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Ellenius, K., Giampiero, S. (2000) The Cost 249 SpeechDat Multilingual Reference Recogniser, Zbornik Second International Conference on Language Recources and Evaluation. Athens, pp. 1351-1355.

Kačič, Z., Horvat, B., Z□gling S. (2000) Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language. Zbornik Second International Conference on Language Recources and Evaluation. Athens, pp. 943-946.

Kačič, Z. (1998) Definiranje leksikona izgovarjav lastnih imen za slovenski jezik. Zbornik konference ERK`98. Portorož, Slovenija, pp. 185-188.

Kaiser., J., Kačič, Z. (1998) Development of Slovenian SpeechDat Database. Proceedings of the Workshop on Speech Database Development for Central and Eastern European Languages. Granada. Spain.

Lindberg, B., Johansen, F.T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Ellenius, K., Giampiero, S. (2000) A Noise Robust Multilingual Reference Recogniser Based on SpeechDat II. Zbornik konference ICSLP, Peking.

Miksić, A., Kačič, Z., Horvat, B. (1998) O-tel – an experimental reverse directory telephone service with barg-in capability. Zbornik Workshop Interactive voice technology for telecommunications Applications, Torino, Italija, pp.171-174.

Rojc, M. Kačič, Z. (1999) Večjezično procesiranje besedil in uporaba na področju procesiranja govora ter naravnega jezika. Zbornik konference ERK`99. Portorož, Slovenija, pp. 201-204.

Rojc, M., Kačič, Z. (2000) A Computational Platform for Development of Morphologic and Phonetic Lexica. Zbornik Second International Conference on Language Recources and Evaluation. Athens, pp. 277-281.

Romih, M. (1998) Amebis in jezikovne tehnologije. Zbornik konference Jezikovne tehnologije za slovenski jezik, Ljubljana, pp. 29-34.

Sepesy Maučec, M. (1998) Modeliranje daljnosežnih korelacij v naravnem jeziku. Zbornik konference ERK`98, pp. 209-212.