# SEDLIS – sistem za avtomatsko identifikacijo jezika iz besedila

**Andrej Žgank, Zdravko Kačič, Bogomir Horvat**

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Smetanova 17, 2000 Maribor, Slovenija
e-mail: andrej.zgank@uni-mb.si

**Povzetek**

V članku bomo predstavili sistem SEDLIS, ki je namenjen identifikaciji jezika iz besedila. Takšni sistemi so lahko primerni za uporabo na področju večjezične sinteze govora. Trenutno delujoči sistem loči med slovenskim, španskim in nemškim jezikom, vendar je možno število podprtih jezikov povečati. Sistem loči jezike na osnovi pogostosti pojavljanja črk v posameznem jeziku. Pri tem so uporabljeni trigramski jezikovni modeli narejeni z Good-Turingovim glajenjem. Razvito orodje SEDLIS je moč vključiti v sintetizator govora, razvit pa je bil tudi grafični vmesnik, ki omogoča preizkušanje sistema.

**Abstract**

In this article we will present SEDLIS – system for automatic language identification from text. Such systems can be used in the area of multilingual speech synthesis. At the moment, the system is capable of distinguishing between Slovenian, Spanish and German language. The number of supported languages can be increased. Languages are distinguished on the basis of letter frequencies in particular language. Trigram language models built with Good-Turing smoothing were used. The developed tool SEDLIS can be included in speech synthesis system. Also, the graphic user interface, for testing the tool, was developed.

## 1. Uvod

S čedalje hitrejšim razvojem sodobnih tehnologij prihaja do znatnih izboljšav tudi na področju vmesnikov med človekom in strojem. Poudarek je predvsem na izboljšanju uporabnosti in prijaznosti. Z večanjem računske zmogljivosti procesorjev, ki delujejo v elektronskih napravah se govor vedno bolj uveljavlja kot komunikacijsko sredstvo.

Naprave so zmožne komunicirati z uporabnikom s pomočjo razpoznavanja govora, rezultate pa mu posredovati s pomočjo sinteze govora. Posebej zanimivo področje je večjezična sinteza govora. Da bi takšen sistem lahko pravilno deloval, mora vedeti v katerem jeziku je napisano besedilo iz katerega mora tvoriti govorni signal. Pri tem je nujna uporaba sistema za identifikacijo jezika iz besedila. Sistemi za identifikacijo jezika se uporabljajo tudi na mnogih drugih področjih uporabe večjezičnih besedil, kot je na primer iskanje informacij in strojno prevajanje.

Dober primer področja uporabe takšnega večjezičnega sistema za sintezo govora bi bil sistem, ki bi omogočal uporabniku, da bi preko telefona lahko poslušal elektronsko pošto, ki jo je sprejel. Prav tako bi bilo možno takšno sintezo govora uporabiti na področju mobilnih telefonov za sprejemanje SMS sporočil.

Na področju identifikacije jezika iz besedila je bilo razvitih že kar nekaj različnih sistemov, ki razločujejo jezike na osnovi kategorij, ki jih bomo opisali v drugem razdelku. Eden izmed njih, ki podpira zelo veliko število jezikov je sistem TextCat, ki je prosto dostopen na strani http://odur.let.rug.nl/~vannoord/TextCat in je bil izdelan na osnovi članka (Cavnar, 1994).

V članku bomo predstavili sistem SEDLIS (SL-ES-DE Language Identification System), ki je namenjen identifikaciji jezika iz danega besedila. Trenutno zaradi dostopnosti baz podatkov sistem deluje za slovenski, španski in nemški jezik, vendar njegova zgradba omogoča razširitev na večje število jezikov. V drugem razdelku bomo predstavili različne metode za identifikacijo jezika, v tretjem razdelku pa sledi predstavitev zasnove in razvoja sistema. Trenutni rezultati delovanja sistema SEDLIS so predstavljeni v četrtem razdelku, ki mu v petem razdelku sledi zaključek.

## 2. Teoretično ozadje identifikacije jezika

Sistemi za identifikacijo jezika lahko temeljijo na različnih lastnostih jezika, ki so podrobneje predstavljene v literaturi s področja jezikoslovja (Nickel, 1985). Identifikacija jezika je zahteven problem, saj je dokaj težko določiti kategorije po katerih se jeziki razlikujejo med seboj (Morgan, 1991). Tudi govorci, ki tekoče govorijo več sorodnih jezikov težko opišejo po čem ločijo posamezne sorodne jezike. Možne osnovne kategorije, po katerih lahko ločujemo jezike med seboj, so:

- Fonologija: jeziki imajo različen nabor fonemov, vendar je del fonemov mnogokrat skupen več jezikom. Med jeziki je tudi razlika v pogostosti pojavljanja posameznega fonema in njihovih zaporedij.
- Prozodika: naglas, trajanje in višina fonema se razlikuje med jeziki.
- Skladnja: način tvorjenja stavkov je različen. Tudi v primeru, da je ena beseda skupna več jezikom, prihaja do razlik v kontekstu, v katerem beseda nastopi.
- Morfologija: način tvorjenja besed, slovarji besed in koreni besed so različni za posamezni jezik.

Naštete razlike med jeziki in njihove kombinacije služijo današnjim sistemom identifikacije jezika kot osnova za delovanje. Prvi dve kategoriji se več uporabljata v primeru, ko moramo identificirati jezik na osnovi analize govornega signala. Ta način ugotavljanja jezika pride do izraza pri večjezičnem razpoznavanju govora, kjer na primer identifikator jezika najprej identificira jezik in nato aktivira razpoznavalnik govora za ta jezik. Drugi

dve kategoriji ločevanja jezikov sta primernejši za identifikacijo jezika iz besedil.

Cilj sistema, ki je bil razvit, je identifikacija jezika iz danega besedila. Pri tem smo se odločili, da bomo kot osnovo uporabili pogostost pojavljanja zaporedja treh črk (trigram) v besedilu in identifikacijo izvedli s pomočjo trigramskega jezikovnega modela (Jelinek, 1997). Tako lahko sistem razvrstimo v kategorijo, ki jezike razločuje na osnovi morfologije. Podoben pristop je bil uporabljen v (Zissman, 1996), kjer pa so na osnovi pogostosti pojavljanja fonemov ugotavljali jezik izgovorjenih stavkov. Glede na zaporedje treh črk $w_{k-2}$, $w_{k-1}$, $w_k$ verjetnost nastopa črke $w_k$ izračunamo kot relativno frekvenco:

$$P(w_k \mid w_{k-2}, w_{k-1}) = \frac{C(w_{k-2}, w_{k-1}, w_k)}{C(w_{k-2}, w_{k-1})}, \qquad (1)$$

pri tem $C$ predstavlja število bigramov in trigramov za posamično kombinacijo črk v tekstu. Kot vidimo iz enačbe (1) je verjetnost nastopa črke $w_k$ odvisna od prejšnjih dveh črk, ki sta se pojavili.

Glavno področje uporabe jezikovnih modelov je predvsem razpoznavanje govora, kjer jih uporabljamo za ugotavljanje verjetnosti nastopa besed. Vrednosti za jezikovni model izračunamo na naboru teksta, ki ga uporabimo za učenje jezikovnega modela. Pri takšnem pristopu se pojavi težava, saj se lahko v besedilu, za katerega želimo ugotoviti jezik, pojavi zaporedje treh črk, ki ga ni bilo v učnem besedilu. Takšnemu zaporedju treh črk bi bila torej pripisana verjetnost nič. Tej težavi se ne moremo popolnoma izogniti niti z večanjem učnega besedila (Jelinek, 1997), še posebej pri infleksijskih jezikih.

Težavo zaobidemo tako, da s postopkom glajenja pripišemo določeno verjetnost tudi zaporedjem, ki jih ni bilo v učnem besedilu. Pri tem moramo uporabiti drugo besedilo, kot je bilo uporabljeno za učenje. Podrobnosti o različnih možnih postopkih glajenja lahko najdemo v (Jelinek, 1997). V našem sistemu SEDLIS smo uporabili Good – Turingov postopek.

Slabost ugotavljanja jezika posamičnega stavka na osnovi pogostosti črk v stavku, ki jo lahko predvidimo je, da v primeru kratkih stavkov sistem ne dobi dovolj trigramov za dobro oceno jezika. Tudi omenjeni sistem TextCat deluje na podobni osnovi kot naš sistem, vendar jezik identificira na osnovi vsote razdalj med N-grami črk v učnem in testnem profilu N-gramov, ki je bil narejen na osnovi besedila.

## 3. Opis sistema

### 3.1. Uporabljene baze

Razvili smo sistem SEDLIS za identifikacijo slovenskega, španskega in nemškega jezika. Za izdelavo in preverjanje sistema smo uporabili baze SpeechDat(II) (Kaiser in Kačič, 1998) za slovenski, španski in nemški jezik. Te baze so sicer namenjene akustičnemu procesiranju govora, vendar je preverjanje pokazalo, da vsebujejo tudi dovolj veliko besedilo za modeliranje jezika. Različnost v zasnovah večjezičnih baz, ki so bile uporabljane v preteklosti za izgradnjo identifikatorjev jezika, je oteževala primerjavo posameznih sistemov med seboj. Glede na veliko število jezikov (Höge et.al., 1999), za katere je že bila ali je v izdelavi baza SpeechDat, je tako ta težava delno odpravljena.

Ker je bil naš cilj modeliranje pogostosti pojavljanja črk v besedilu, smo iz baze SpeechDat(II) uporabili nabor fonetično uravnoteženih stavkov, ki so bili izbrani s pomočjo skripte projekta COST249 "SpeechDat Task Force" (Johansen et.al., 2000). Ker je baza SpeechDat(II) namenjena predvsem razvoju sistemov avtomatskega razpoznavanja govora, je njena sestava takšna, da več govorcev prebere isti stavek. Da bi izločili vpliv teh ponovitev stavkov v učnem besedilu na izračun jezikovnega modela, smo celotno učno besedilo preuredili tako, da se vsak stavek v njem pojavi samo enkrat. Prav tako smo iz besedila izločili vse velike črke, saj bi to poslabšalo modeliranje jezika. Osnovni podatki o uporabljenem naboru stavkov so podani v tabeli 1.

| | Jezik | | |
|---|---|---|---|
| Št. črk | SL | ES | DE |
| Učenje | 24.000 | 153.000 | 172.000 |
| Prever. | 865 | 1.200 | 1.800 |
| Slovar | 28 | 31 | 34 |

Tabela 1: Podatki o stavkih uporabljenih za učenje in preverjanje (slovenski (SL), španski (ES) in nemški (DE) jezik).

V vsakem stavku smo uporabili posebna simbola, ki označujeta začetek in konec stavka. Pri izdelavi jezikovnega modela smo kot posebno črko upoštevali tudi presledek med dvema besedama. S tem smo v jezikovnem modelu upoštevali tudi način tvorjenja stavkov v posamičnem jeziku. Kot vidimo v tabeli 1 je slovar črk največji za nemški jezik (34 znakov) in najmanjši za slovenski jezik (28 znakov). Vpliv črk, ki so značilne za posamičen jezik (npr.: "č" v slovenščini) ni posebej upoštevan, saj se lahko zgodi, da beseda s takšno črko nastopi tudi v stavku v drugem jeziku, na primer kot lastno ime. Stavki uporabljeni za preverjanje delovanja sistema so bili naključno izbrani iz nabora fonetično uravnoteženih stavkov v bazi SpeechDat(II) in se niso nahajali v učnem besedilu. Trenutno je sistem sposoben razlikovati med tremi jeziki, vendar je moč njihovo število, v primeru dostopnosti večjega števila baz, tudi povečati.

### 3.2. Zgradba in delovanje sistema

Celotni sistem smo razvili na delovnih postajah HP-UNIX, vendar ga je možno z manjšimi prilagoditvami prenesti v okolje Windows. Za izdelavo treh potrebnih jezikovnih modelov smo uporabili standardno orodje za gradnjo jezikovnih modelov CMU-SLM Toolkit V2.0 (Clarkson, 1997) univerz Carnegie-Mellon in Cambridge, ki je prosto dostopno na naslovu http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html. Druga potrebna orodja za delovanje sistema smo razvili v programskem jeziku ANSI C, ki omogoča hitro delovanje sistema.

Sistemu smo za boljšo predstavitev delovanja dodali grafični vmesnik, razvit v skriptnem jeziku Tk.

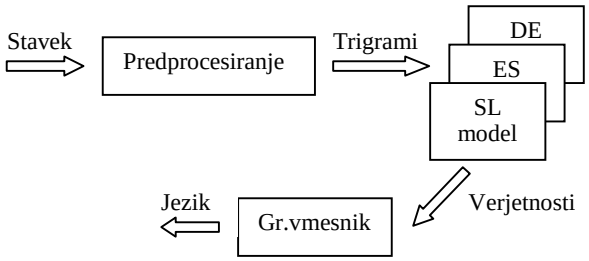
Na urejenem in preverjenem učnem besedilu smo najprej s pomočjo orodja CMU-SLM zgradili trigramske jezikovne modele. Pri tem smo uporabili Good - Turingovo glajenje z umikanjem. Podatke o dobljenih trigramskih jezikovnih modelih podaja tabela 2.

| Jezik | Število | | | Perpl. |
|---|---|---|---|---|
| | 1g | 2g | 3g | |
| SL | 28 | 506 | 3159 | 9,53 |
| ES | 35 | 584 | 4270 | 7,05 |
| DE | 38 | 817 | 6753 | 7,15 |

Tabela 2: Število unigramov, bigramov in trigramov za posamični jezik, izračunana je tudi perpleksnost modela.

Najmanjše število bigramov in trigramov je v slovenskem jeziku, kar je posledica dejstva, da je v slovenskem jeziku tudi najmanj črk. Perpleksnost posameznega jezikovnega modela smo izračunali z 20 testnimi stavki za vsak jezik. Glede na dejstvo, da je perpleksnost jezikovnega modela za slovenski jezik za eno tretjino večja od perpleksnosti ostalih dveh jezikovnih modelov, je možno na osnovi izkušenj iz uporabe jezikovnih modelov pri razpoznavanju govora predpostaviti, da bo rezultat identifikacije jezika za slovenski jezik najslabši. Stavke uporabljene za izračun perpleksnosti jezikovnih modelov smo kasneje uporabili tudi za testiranje uspešnosti identifikacije jezika.

Tako pripravljene jezikovne modele smo vključili v sistem. Zgradbo sistema na najvišjem nivoju prikazuje slika 1:




Slika 1: Prikaz delovanja sistema SEDLIS za identifikacijo jezika na najvišjem nivoju.

Del za predprocesiranje izloči iz vhodnega besedila stavek in ga spremeni v obliko, ki jo uporablja sistem, tako da stavku doda oznako za začetek in konec stavka, ter vse presledke spremeni v ustrezni model. Zaporedje trigramov črk potuje v vse tri trigramske jezikovne modele, ki delujejo vzporedno. Za vsak trigram iz stavka v jezikovnem modelu poiščemo verjetnost tega trigrama, ki jo prištejemo k verjetnostim prejšnjih trigramov iz stavka. Jezik jezikovnega modela, ki je dal največjo verjetnost je prepoznan kot pravilni rezultat. Takšna zasnova sistema omogoča pravilno identifikacijo jezika za posamične stavke tudi v primeru, ko je v enem besedilu uporabljenih več različnih jezikov.

Za lažjo in boljšo komunikacijo s sistemom smo razvili tudi grafični vmesnik, ki ga kaže slika 2.


Slika 2: Izgled grafičnega vmesnika. Sistem SEDLIS je pravilno prepoznal, da je vnešeni stavek v slovenskem jeziku.

Grafični vmesnik omogoča prepoznavanje jezika samo za posamični stavek. Pravilen rezultat je prikazan tako, da se osvetli jezik, v katerem je vnešeni stavek, kot lahko vidimo na sliki 2. Sistem je pravilno prepoznal, da je stavek “Kateri jezik je to?” zapisan v slovenskem jeziku.

## 4. Rezultati

Na začetku razvoja sistema SEDLIS smo iz učne baze izločili dva nabora stavkov, ki smo jih uporabili za preverjanje delovanja sistema. V prvem naboru stavkov je bilo 20 stavkov za vsak posamezni jezik. Rezultate identifikacije jezika s prvim testnim naborom stavkov podaja tabela 3.

| | Jezik | | |
|---|---|---|---|
| | SL | ES | DE |
| Ident. jezika | 90% | 95% | 100% |

Tabela 3: Rezultati delovanja sistema SEDLIS za razpoznavanje slovenskega (SL), španskega (ES) in nemškega (DE) jezika.

Kot vidimo v tabeli 3 je sistem najboljše rezultate dosegel pri identifikaciji nemškega in najslabše pri identifikaciji slovenskega jezika. Na osnovi podatkov o velikosti učnega besedila v tabeli 1, je možno sklepati, da je to posledica različno velikih učnih besedil. Do te razlike je prišlo zaradi razlik med bazami SpeechDat(II) in zaradi uporabljene skripte projekta COST249 (Johansen et.al., 2000), ki je iz učnega besedila izločila tiste stavke, pri katerih so posnetki slabe kvalitete.

Pri preverjanju, v katerih primerih je sistem napačno prepoznal jezik, se je izkazalo, da se je to zgodilo pri zelo kratkih stavkih, kot smo tudi predpostavili. To je posledica dejstva, da iz takšnih stavkov jezikovni model dobi majhno število trigramov, na osnovi katerih lahko poda oceno o jeziku, v katerem je vnešeni stavek. V primeru, da bi uspeli izboljšati delovanje sistema SEDLIS tudi za kratke stavke (posamezne besede), ga bi bilo možno uporabiti tudi znotraj enojezičnega slovenskega sintetizatorja govora. V takšnem sistemu bi SEDLIS na nivoju analize besedila preverjal jezik posamezne besede. V primeru, ko se znotraj slovenskega stavka nahaja beseda v tujem jeziku, na primer lastno ime, se lahko za takšno besedo uporabijo drugačna pravila tvorjenja fonetičnega zapisa.

V drugem testnem naboru stavkov je bilo 153 različnih stavkov v slovenskem jeziku. S temi stavki smo naredili

identifikacijo jezika in preverili kakšno je razmerje med različnimi napačno identificiranimi stavki. Te rezultate podaja tabela 4.

| | Ident. jezika | | |
|---|---|---|---|
| | SL | ES | DE |
| SL stavki | 95,43% | 1,96% | 2,61% |

Tabela 4: Rezultati identifikacije 153 stavkov v slovenskem jeziku (slovenski (SL), španski (ES) in nemški (DE) jezik).

V primeru ko smo za identifikacijo jezika uporabili drugi testni nabor stavkov, smo dobili nekoliko boljše rezultate za identifikacijo slovenskega jezika, kot pri prvem naboru. To je verjetno posledica dejstva, da smo imeli v drugem primeru na razpolago večje število stavkov in se tako napake niso tako močno poznale pri skupnem rezultatu. Kot vidimo v tabeli 4 je razmerje napačne identifikacije slovenskega jezika približno enakomerno porazdeljeno na španski in nemški jezik. Iz tega lahko sklepamo, da so modeli za vse tri jezike dovolj dobro naučeni.

Uspešnost sistema bi bilo moč izboljšati tako, da bi povečali število stavkov v bazah besedil, ki smo jih uporabili za učenje jezikovnih modelov sistema SEDLIS. To še posebej velja za slovenski jezik, saj je bila velikost učne baze zanj samo ena petina velikosti ostalih dveh učnih baz. Možno je, da bi z drugo metodo glajenja trigramskih jezikovnih modelov, lahko dosegli še nekoliko boljše rezultate, saj se Good – Turingovo glajenje v določenih primerih premajhnega učnega besedila (Jelinek, 1997) ne obnese najbolje.

## 5. Zaključek

Na osnovi dobljenih rezultatov lahko zapišemo, da je sistem primeren za uporabo v realnih aplikacijah s področja jezikovnih tehnologij, predvsem večjezične sinteze govora. V prihodnje načrtujemo razširitev sistema na večje število jezikov (angleški jezik), vendar bo pri tem potrebno upoštevati tudi ali je jezik s stališča sinteze govora sploh smiselno vključiti. Celotni sistem bo prenešen v HTML okolje na medmrežje, kjer bo javno dostopen za preizkušanje.

## *Zahvala*

Za dovoljenje uporabe nemške in španske baze SpeechDat(II) se avtorji zahvaljujejo podjetju Siemens AG in univerzi Universitat Politechnica de Catalunya.

## 6. Literatura

Cavnar W.B., J.M. Trenkle, 1994, N-Gram-Based Text Categorization. *Proc. of Third Annual Symposium on Document Analysis and Information Retrieval.* Las Vegas, USA.

Clarkson P.R., R. Rosenfeld., 1997, Statistical Language Modeling Using the CMU-Cambridge Toolkit. *Proc. Europ. Conf. Speech Proc. and Techn. (EUROSPEECH).*

Höge, H., C. Draxler, H. van den Heuvel, F.T. Johansen, E. Sanders, H.S. Tropf, 1999, SpeechDat multilingual speech databases for teleservices: Across the finish line. *Proc. Europ. Conf. Speech Proc. and Techn. (EUROSPEECH).*

Jelinek, Frederic, 1997, *Statistical Methods for Speech Recognition*, MIT Press.

Johansen, F.T., N. Warakagoda, B. Lindberg, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, G. Salvi, 2000, The COST 249 SpeechDat multilingual reference recogniser, *Second international conference on language resources and evaluation.* Athens, Greece.

Kaiser, J., Z. Kačič, 1998, Development of the Slovenian SpeechDat database, *First International Conference on Language Resources and Evaluation*. Granada, Spain.

Morgan, D.P., C.L. Scofield, 1991, *Neural Networks and Speech Processing*, Kluwer Academic Publishers.

Nickel Gerhard, 1985, *Einführung in die Linguistik: Entwicklung, Probleme, Methoden*, Berlin.

Zissman, M.A., 1996, Comparison of Four Approaches to Automatic Language Identification of Telephone Speech, *IEEE Trans. on Speech and Audio Processing*, 1:41-44.