Jezični alati za obradu njemačkog jezika: distribucija slogova u njemačkom jeziku

Babić, Nikola

SVEUČILIŠTE U RIJECI
ODJEL ZA INFORMATIKU
Preddiplomski dvopredmetni studij informatike

JEZIČNI ALATI ZA OBRADU NJEMAČKOG JEZIKA: DISTRIBUCIJA SLOGOVA U NJEMAČKOM JEZIKU

Završni rad

Autor: Nikola Babić
Mentor: Izv. prof. dr. sc. Sanda Martinčić-Ipšić

U Rijeci, Rujan 2015.

JEZIČNI ALATI ZA OBRADU NJEMAČKOG JEZIKA: DISTRIBUCIJA SLOGOVA U NJEMAČKOG JEZIKU

SAŽETAK

Jezični alati služe za detaljnu računalnu analizu prirodnog jezika, čije rezultate možemo iskoristiti u raznim istraživanjima. Moguća je analiza na temelju točno određenih elemenata koji se nalaze u prirodnom jeziku. Takvi alati su poznati pod nazivom „Natural Language Processing Tools“. Svrha ovog rada je prikazivanje aspekta rastavljanja slogova vezano uz pravila njemačkog jezika i provjera dobivenih rezultata istraživanja. Za provođenje statističke analize slogova korišten je korpus koji sadrži riječi iz pojednostavljenog njemačkog rječnika. Konačni rezultat programske analize se uspoređuje s postojećim rastavljenim slogovima riječi unutar njemačkog rječnika, što ukazuje na određene pogreške u slučaju prepuštanja programskog rastavljanja slogova. Rastavljanje na slogove je važan korak u problemu određivanja naglasaka riječi pri umjetnoj tvorbi govora. Boljim razumijevanjem prirodnog jezika smo korak bliže boljem razumijevanju nas samih, pošto je jezik proizvod ljudski bića.

Uz pomoć programa za automatsko rastavljanje riječi na slogove bila je omogućena statistička analiza pojedinih detalja slogova unutar njemačkog jezika. Uspješno je analizirana učestalost broja slogova unutar svake riječi, duljina svakog pojedinog sloga i frekvencija pojavljivanja 16 modela sloga u svakoj riječi. Na kraju analize rezultata je ustanovljeno:

- *da su slogovi duljine tri slova najučestaliji (46%), te da slogovi većih duljina nisu zastupljeni u tolikoj mjeri (ukupno oko 20%);*
- *da ima najviše riječi koje sadrže po četiri i pet različitih slogova (oko 40%);*
- *da je najveći broj različitih slogova među slogovima sastavljenih od 5 riječi (oko 25%)*
- da je „CVC“ najzastupljeniji model sloga (44%) što odmah potvrđuje rezultate o slogovima duljine tri slova.

Ključne riječi: slog, distribucija slogova, naglasak, izgovor, fonem, morfem

Language Processing Tools for German Language: The Distribution of Syllables in German Language

ABSTRACT

Language processing tools are commonly used in researches for analyzing and thus better understanding of the natural language. The German language is known for its many rules in language structure and is therefore a good example for analyzing. Syllables are one of the most important units of a language, because it shows the structure of every word in the language. Even though syllables are that important, they are not the main focus in research among the linguistic scientists.

The goal is to test the algorithm for automatic syllabification in the German language from the main corpora of dictionary words. The implementation of German grammar rule into the program is also a high priority.

The results of the automatic syllabification are split into different focus areas of syllables in language: list of all possible syllabic models, the frequencies of syllables according to the number of letter and the frequencies of syllables within each word. The analysis of the different focus areas gave a definitive answer to the questions. CVC model within the German language are by far the most frequent (44%). Syllables that are made out of three letters are the most frequent kind of syllables (46%) while there are other syllables with more letters are considerably fewer. The most frequent number of syllables within a word is five syllables per word. The last assignment was to the algorithm and to evaluate it (with approximated error that is below 40%).

KEY WORDS: syllable, distribution of syllables, accent, pronunciation, phoneme, morpheme

Sprachverarbeitungswerkzeuge für die deutsche Sprache: Silbendistribution in deutscher Sprache

ZUSAMMENFASSUNG

Sprachverarbeitungswerkzeuge werden häufig in Untersuchungen benutzt für das Analysieren und besseres Versťandnis der natürlichen Sprache. Die deutsche Sprache ist für bekannt für ihre viele Regeln in der Sprachstruktur und ist damit ein gutes Beispiel für eine solche Analyse. Silben sind eine der wichtigsten Einheiten der Sprache, weil sie die Struktur jedes Wortes in der Sprache zeigen. Obwohl Silben solch eine wichtige Rolle haben, sind sie nicht die Hauptproblematik der Sprachwissenschaftlern.

Das Ziel ist es, den Algorithmus für die automatische Silbentrennung in der deutschen Sprache zu testieren mit Wörtern aus dem Wörterbuch. Auch eine hohe Priorität hat die Durchführung der Regeln aus der deutschen Grammatik in das Programm.

Die Ergebnisse der automatischen Silbentrennung sind in verschiedene Schwerpunktbereiche der Silben aufgeteilt: Liste aller möglichen Silbenmodelle, die Häufigkeit von Silben an Hand von der Anzahl der Buchstaben und Häufigkeit verschiedener Silben in jedem Wort. Die Analyse der verschiedener Schwerpunkten gab endgültige Antworten auf die Fragen. CVC-Modell sind mit Abstand am häufigsten (44%). Die häufigste Art von Silben an Hand von Buchstaben sind Silben, die aus drei Buchstaben strukturiert sind (46%), während es deutlich weniger andere Silben gibt. Die häufigste Anzahl von Silben innerhalb eines Wortes ist fünt Silben pro Wort. Die letzte Aufgabe war das Bewerten der Genauigket des Algorithmus im Programm (mit ungefähr 40% Fehler).

Schlüsselwörter: Silbe, Silbendistribution, Akzent, Aussprache, Phoneme, Morpheme

1. UVOD U NLP NJEMAČKOG JEZIKA

U širem smislu je lingvistika znanstvena disciplina koja se bavi istraživanjem i objašnjavanjem prirodnog i govornog jezika. Bio je potreban dugi niz stoljeća dok lingvistika nije postala priznata kao ravnopravna znanstvena disciplina. Već su prvi filozofi utvrdili da jezik igra glavnu ulogu razlikovanja među ljudima i životinjama. Tijekom istraživanja utvrđene su tri svrhe, odnosno aspekta, jezika: antropološki, aspekt unutarnje jezične strukture i komunikativno-socijalni aspekt. Oni upućuju na govornu sposobnost čovjeka, shvaćanje sistematske složenosti jezika i korištenje jezika u svrhu djelovanja.

Postepeni razvoj jezika i tehnologije omogućili su stvaranje računalne lingvistike. Danas je moguće uz pomoć računala provesti analizu teksta i u kratkom vremenu dobiti precizne rezultate. NLP, odnosno „*Natural Language Processing*“, je posebno područje računarstva, umjetne inteligencije i računalne lingvistike koji se fokusira na automatsku obradu prirodnog jezika računalima [Bird, Klein i Loper, 2009: 1-3]. Zadaci računalne analize jezika mogu biti jednostavni poput brojanja učestalosti određenih vrsta riječi, dok s druge strane mogu izvoditi kompliciranu analizu poput razumijevanja ljudskih iskaza. Zbog naglog širenja takvih sustava danas se pogotovo vidi utjecaj na mobilnim uređajima, gdje igra veliku ulogu u višejezičnim aplikacijama. Alati koji nam dopuštaju izvođenje tih zadataka zovu se „*Natural Language Processing Tools*“.

Razlaganje riječi na nerazložne jedinice, odnosno slogove, ne predstavlja intenzivno interesno područje istraživanja lingvista njemačkog govornog područja. Razlog tome je jednostavnost vlastite intuitivne procjene kada i gdje je potrebno razlagati riječ prema određenim pravilima. Problem se javlja pri implementaciji svih pravila razlaganja u jezik razumljiv računalu uz sve iznimke prirodnog jezika [Gojmerac, 1992: 1].

NLP se također bavi problemima automatske analize teksta na semantičkoj razini, odnosno zaključivanja biti iz ulaznih tekstova. Takva analiza teksta je moguća zahvaljujući njemačkom modelu „*Saarbrücker Pipelinemodell*“ [Tapken, 2013].
Saarbrücker model se provodi kroz 5 različitih koraka koji uključuje:

- Prepoznavanje tokena – lanac slova se segmentira u jezične jedinice (riječi i rečenice);

- Morfološku analizu – analiziraju se osobni oblici i padežne oznake kako bi se izvukle gramatičke informacije, te riječi iz teksta vratile u osnovne oblike;
- Sintaksnu analizu – riječi iz teksta se analiziraju prema njihovoj strukturalnoj funkciji (subjekt, objekt, broj, itd.);
- Semantičku analizu – Rečenicama, odnosno njihovim dijelovima, se dodjeljuje značenje. Ovaj korak potencijalno sadrži mnogo različitih pojedinačnih koraka, pošto se značenje riječi teško određuje;
- Analizu razgovora i diskursa – Utvrđuje se poveznica između rečenica koje naknadno slijede jedna nakon druge (npr. poveznica između pitanja i odgovora ili objašnjenje činjenicama nakon iznošenja izjave).

Cilj ovog istraživanja je rastavljanje pojedinačnih riječi iz rječnika njemačkog jezika na zasebne slogove uz pomoć automatske distribucije slogova, koja se temelji na gramatičkim, fonetskim i fonološkim pravilima njemačkog jezika. U svrhe istraživanja će se koristiti korak iz prethodno prikazanog modela za automatsku analizu teksta. Morfološka analiza pomaže pri rastavljanju složenih riječi u njihov najosnovniji oblik, što uključuje slogove. Konačni rezultat je vidljiv kao učestalost rastavljenih riječi pri distribuciji slogova. Zbog mogućih pogrešaka tijekom automatske analize i rastavljanja riječi, provedena je dodatna ručna provjera koja služi kao usporedba automatskog i ručnog načina rastavljanja riječi na slogove.

U idućem poglavlju je objašnjenje morfološke analize riječi njemačkog jezika. U to je uključeno objašnjenje značaja fonetike u određivanju slogova. Treće poglavlje prikazuje sva pravila gramatičkog i fonološkog rastavljanja riječi na slogove i uključuje iznimke od tih pravila. U četvrtom poglavlju je objašnjenje same implementacije tih pravila pri automatskom rastavljanju na slogove u programskome jeziku Python, dok je u petom poglavlju prikazana analiza dobivenih rezultata.

2. MORFOLOŠKA ANALIZA

Slogovi u njemačkom jeziku tvore glasovnu (fonetsku) cjelinu. Slog je gramatički, odnosno lingvistički pojam, koji tvori cjelinu iz jednog ili više uzastopnih glasova, tj. fonema, koji se mogu izgovoriti i time dokazati govornu tvorevinu. Usto slogovi čine najmanju glasovnu grupu prirodnog jezika. To znači da je slog dio svake riječi, bez obzira na broj slogova koji ga čine. Pošto je zadatak istraživanja rastaviti slogove iz riječi, potrebna je morfološka analiza kako bi se klasificirali slogovi. Za određivanje punog značaja slogova i morfološke analize riječi, potrebno je uključiti razvoj lingvistike kao znanosti.

Za veliki razvoj lingvistike kao znanstvene discipline bio je zaslužan švicarski lingvist Ferdinand de Saussure [Gojmerac, 1992: 20-21]. Bavio se istraživanjima i osmišljavanju teorija jezika na indogermanskim govornim područjima. Zbog njega se počinju pokretati strukturalističke škole, koje su skupa sa praškim školama vršile veliki utjecaj na razvoj novih teorija jezika. De Saussure je sistematizirao jezik pod 3 pojma: *langue (*„jezik“), *parole* („govor“), *language* („ljudski govor“ što ujedinjuje *langue* i *parole*). *Langue* označava sistem pojedinačnog razumijevanja jezika, dok *parole* ukazuje na konkretnu uporabu jezika pri govoru i pisanju. Također bitno za napomenuti je njegov drugi sistem u kojem spaja jezik i jezične znakove. De Saussure ukazuje na nedvojbenu povezanost onog što se određuje i slikovne projekcije toga što se određuje u mislima, tj. zamisli. „*Zamisao je prednja strana dok je glas stražnja strana; prednja strana se ne može prerezati bez da se istovremeno prereže i stražnja strana.*“ [Gojmerac, 1992: 23]

Tek početkom praške škole, koja se smatra prvim pravim počecima znanosti o jeziku, su se počele razvijati nove teorije lingvistike prema jezičnim konceptima de Saussurea [Gojmerac, 1992: 32]. Zastupnici praške škole su označili svoju lingvistiku kao funkcionalnu i strukturiranu. Među zastupnicima te škole su bili lingvisti N.S. Trubetzkoy i R. Jakobson. Centralna lingvistička disciplina praške škole je fonologija, koju je utemeljio Nikolai Sergejewitsch Trubetzkov. Prema Trubetzkoyu, tvorba jezika i čin govora oba su nerazdvojna dijela fenomena zvan jezik. On definira fonetiku kao znanost o materijalnoj strani ljudskog govora. Fonetika ne istražuje fizičku kvalitetu glasa, već samo njenu funkciju u određenom jezičnom sistemu, pogotovo njena funkcija diferenciranja značenja riječi. Tu glasovnu promjenu Trubetzkoy naziva fonem. Svaka riječ mora imati točno određen broj fonema u točno određenom

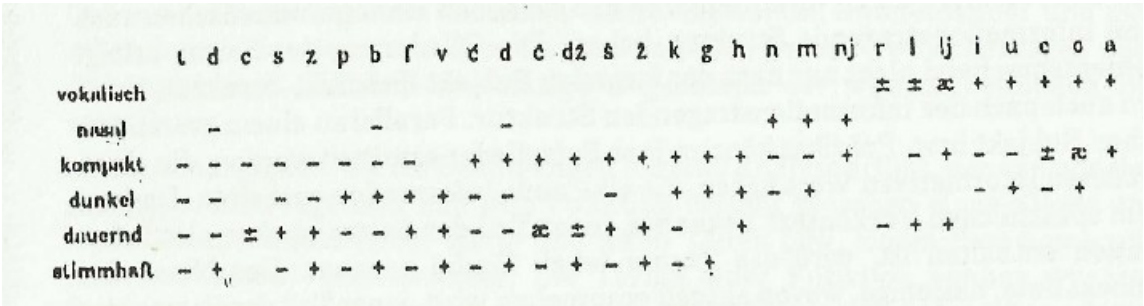
redosljedu kako bi se ukazala razlika od ostalih riječi u jeziku. Fonem sam po sebi nema nikakvo značenje, niti je najmanja jedinica koja sadrži značenje, već služi kao nešto po čemu će se riječ kasnije razlikovati od ostalih.

Postoji skoro neograničen broj glasova u jeziku koje čovjek može izgovoriti, ali postoji samo određeni broj glasova koji će utjecati na promjenu značenja kao što su fonemi. Na primjeru sljedećih hrvatskih i njemačkih riječi vidljivi su fonemi:

- njem. Reise – leise; Bitte – Sitte,
- hrv. presti – plesti; blijed – slijed.

Ovakve riječi tvore tzv. “minimalni par”, jer su fonemi između nasuprotnih riječi uzrokovali najmanju promijenu u glasu i pritom promijenili cijelo značenje riječi.

Unutar praške škole je u razvoju fonologije bio jednako bitan i Roman Jakobson. Za njega fonemi predstavljaju skupinu karakterističnih obilježja, koja su odabrana iz univerzalno valjanog inventara obilježaja. Jakobson dijeli obilježja na nerazdvojiva i prozodijska. Prozodijska obilježja se dijele na ton, jačinu, i količinu. Koriste se u opoziciju sa drugim riječima kako bi poslužile kao dokaz promjene (npr. Haus – Maus; /h/ postaje /m/). Nerazdvojiva obilježja je Jakobson podijelio na 12 binarnih opozicija koje vrijede univerzalno. Prvih 9 čine zvukovne značajke (npr. vokalni – ne vokalni; konzonantni – ne konzonantni) dok zadnje 3 čine tonalitetne značajke (npr. dubok – ne dubok; oštro - ne oštro). Ove opozicije su dovoljne za fonološko opisivanje svakog pojedinog jezika kao što je prikazano za hrvatski jezik na slici 1 [Gojmerac, 1992: 37].



Slika 1: Podjela 12 binarnih opozicija u hrvatskom jeziku prema Jakobsonu [Gojmerac 1992]

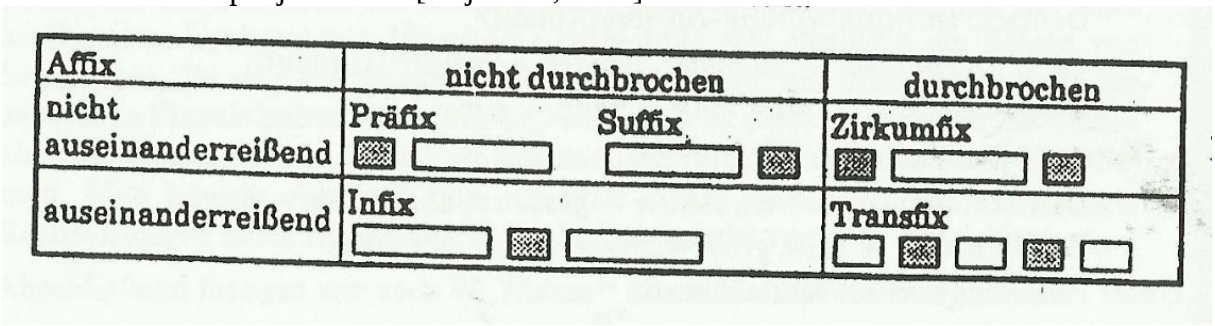
Glavni jedinica morfologije je riječ [Gojmerac, 1992: 147]. Morfologiju možemo opisati kao znanstvenu disciplinu koja se bavi istraživanjem riječi. Dakle, nalazi se između disciplina koje se bave istraživanju glasova i rečenica. No riječ nije čvrsto definirana jedinica, stoga se pojam „riječ“ može objasniti samo kroz različite indikatore. Među njih spadaju ortografski, fonološki, gramatički, distribucijski i semantički indikatori. Kako riječ ne može služiti kao glavna jedinica, morfologija koristi morfem kao jedinicu za razmatranje konstrukcija riječi. Morfem je najmanji jezični znak, koji se ne može rastaviti na jednostavnije jedinice. Spajanjem različitih morfema dobivamo složene riječi. U primjeru sljedeće rečenice je vidljivo rastavljanje na morfeme:

> /Ein/ klein/ er/ Jung/ e/ zer/ brach/ das/ Küche/ n/ fenster/ mit/ ein/ em/ leder/ n/ en/ Ball/

Ovu rečenicu je moguće podijeliti na 18 zasebnih jedinica, točnije morfema. Pojedini morfemi mogu biti istovremeno i zasebne riječi (npr. /ein/, /das/, /Ball/), dok pojedini morfemi nikada ne smiju sami stajati u rečenici (npr. /er/, /e/, /zer/, /n/, /em/, /en/). Prema tome se morfemi mogu podijeliti u dvije klase: slobodne i povezane morfeme. Povezani morfemi se drugim imenom zovu afiksi. Povezani morfemi imaju gramatičko značenje, mogu služiti za derivaciju novih riječi ili omogućuju kombinaciju (kompoziciju) od najmanje 2 slobodna morfema. Među slobodnim morfemima postoje i leksemi, što su morfemi koji se mogu pronaći u rječniku kao zasebne riječi. Oni morfemi koji pak ovise o kontekstu rečenice se nazivaju deiktički morfemi (npr. ich, er, dieser, dies, itd.).

Zbog mogućnosti sintetiziranja velikog broja riječi u njemačkom jeziku samim nadovezivanjem zasebnih morfema, postoje i različita pravila nadovezivanja. U njemačkom jeziku je izrazito jednostavno tvoriti složenice iz jednostavnih riječi. Za takvo stvaranje riječi su potrebna 2 ili više osnovna slobodna morfema koji se spajaju znakom /s/ (npr. Alter-tum + s + kunde, Heiter-keit + s + aus-bruch). Afiksi služe za nadovezivanje na slobodne morfeme i time upotpuniti ili promijeniti značenje. Oni se mogu dodati na početak (prefiks), kraj (sufiks) ili početak i kraj slobodnog morfema (npr. Be-schreiben, Leit-ung, Ge-brüll-e). Primjeri načina podjele afiksa su grafički predočeni u tabeli 1.

Tabela 1: Načini podjele afiksa [Gojmerac, 1992]



3 RASTAVLJANJE NA SLOGOVE

Iz prethodne cjeline smo utvrdili da je slog čvrsto povezan uz izgovor, odnosno glasovni slijed koji čovjek proizvodi. Također je glavni fonetski i fonološki element svake riječi. Kako je materinski jezik usađen svakom govorniku čitavom strukturom zvuka i gramatike, tako je za govornike tog jezika intuitivno jasno rastavljanje riječi na pojedinačne slogove koji tvore tu riječ. Za bolje razumijevanje metoda rastavljanja riječi na slogove potrebno je navesti dvije teorije koje spadaju pod nelinearnu fonologiju: autosegmentalna fonologija i metrička fonologija.

3.1 AUTOSEGMENTALNA FONOLOGIJA

Intonacija, ton i naglasak su apstraktni pojmovi koji se ne mogu adekvatno objasniti preko linearne fonologije, zbog čega jer nastala nelinearna fonologija. Problem nastaje pri istovremenom dodjeljivanju pojedinih fenomena, odnosno slogova, većem broju segmenta pojedine riječi [Gojmerac, 1992: 81]. Ako promatramo riječi *Vater* uočavamo kako je sastavljena od dva sloga: *Va-ter*. između razine riječi i sloga se dodaje nova razina zvana „CV-razina“, kako bi se svaki slog dodjelio određenom segmentu. Ta razina ima sličnosti sa skraćenim oznakama za pojmove konsonant i vokal, no zapravo ukazuje na apstraktniju poveznicu. Preko ove razine je objašnjena teorija zvukova jer su to zapravo „vremenske cjeline“. V-cjelina služi kao jedinica zvukovnog prijenosa. U primjeru je dugi vokal predstavljen dvijema V-cjelinama. Slogovi se označavaju ako „autosegmenti“, pošto slogovi posjeduju status poput samostalnih segmenata.

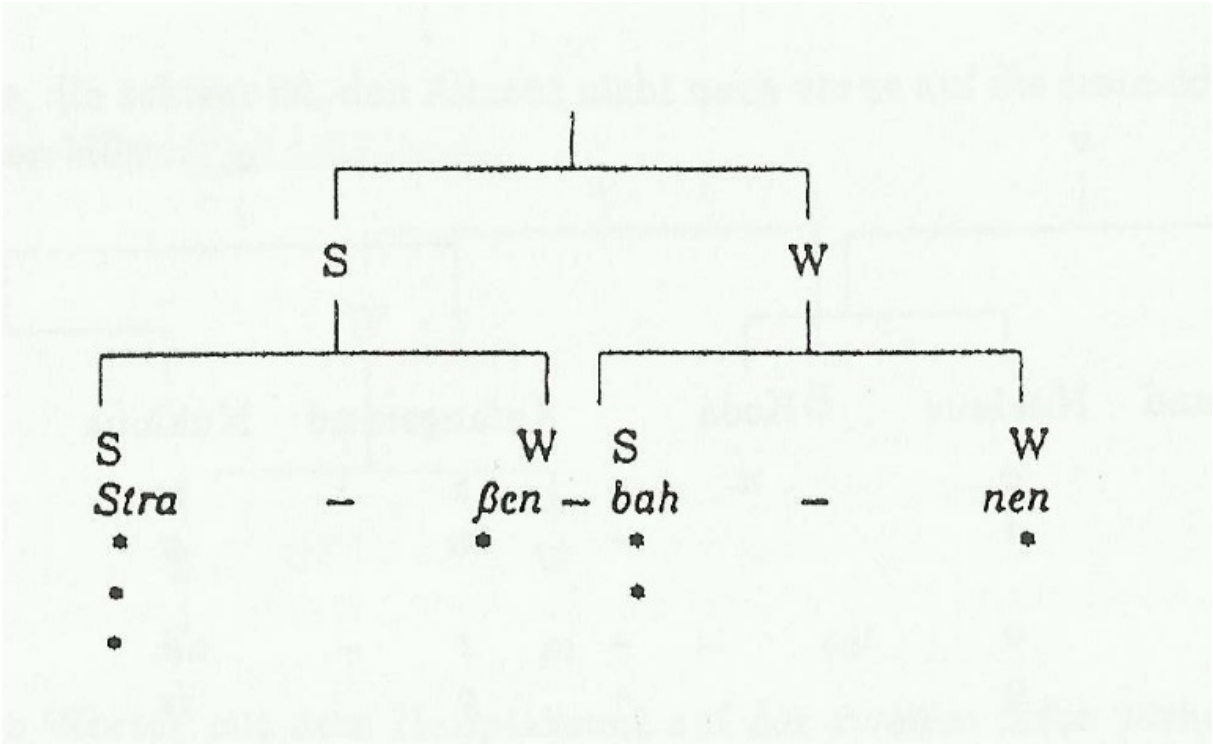
3.2 METRIČKA FONOLOGIJA

Metrička fonologija se bavi istraživanjem poveznica između naglasaka jezika. Svaka riječ koja se sastoji od najmanje dva sloga ima slog koji se ističe glasovno jače od drugog. Ova pojava se može prikazati na različite načine:

- pri većoj glasovnoj jačini naglašenijeg sloga,
- pri dužem trajanju naglašenijeg sloga,
- pri povećanju i sniženju visine tona naglašenijeg sloga.

Glavni cilj metričke fonologije je opisati poveznice naglašavanja u izgovoru. izgovori se zatim rastavljaju na strukturirane dijelove koji se analiziraju na poseban način. Pojedini strukturno bitni dijelovi se označavaju sa *s* („strong“) i *w* („weak“), gdje *s*

označava naglašeniji slog od onog sloga koji je označen sa *w* [Gojmerac, 1992: 82-83].



Slika 2: Ritmička struktura naglaska u slogovima [Gojmerac, 1992]

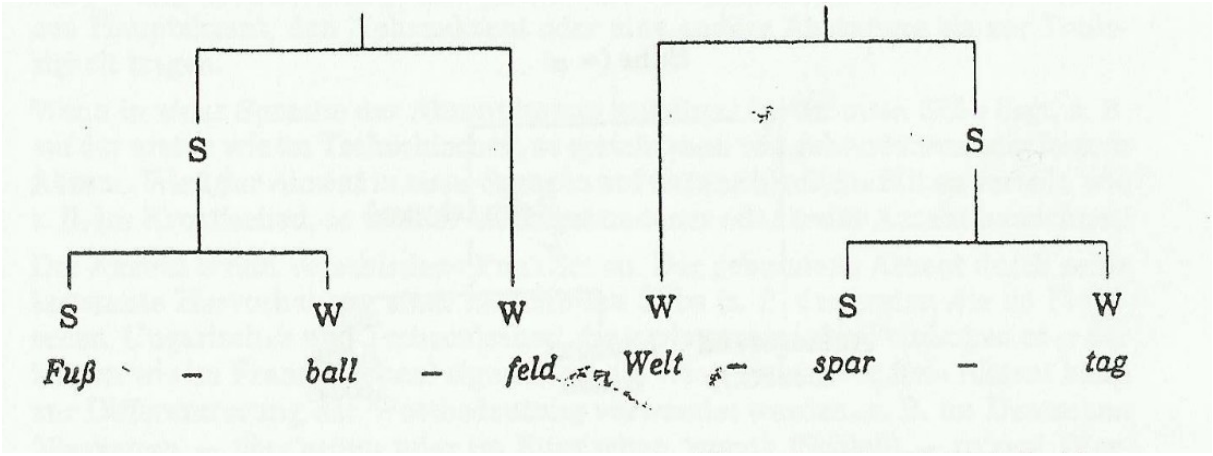
U prikazanom primjeru na slici 2 za riječ *Straßenbahnen* su slogovi *„Stra“* i *„bah“* naglašeniji slogovi od slogova *„ßen“* i *„nen“*. Na višoj razini uočavamo kako je cijeli slog *„Straßen“* naglašeniji od sloga *„bahnen“*. No na ovaj način nije moguće vidjeti logičnu poveznicu zbog čega je slog *„Stra“* naglašeniji od sloga *„ßen“*. Zato se uvodi pojam „metričke rešetke“ koje dodaje poseban znak („*“) svakom slogu i time vizualno prikaže naglašenije slogove svake riječi.

3.3 PRAVILA NAGLASKA U NJEMAČKOM JEZIKU

Pojedine poveznice između naglasaka su toliko kompliciranе u njemačkom jeziku, da ih je teško izraziti općenitim pravilima. Stoga, ćemo se ograničiti na određivanje naglaska u složenim riječima koje su sastavljene od dva i tri sloga [Gojmerac, 1992: 84-85]. Za bolje objašnjenje ovih pravila mogu se koristiti metričke građe i metričke rešetke. Unutar strukture složenih riječi sastavljenih od tri sloga mogu se često uočiti dvije vrste strukture stabla.

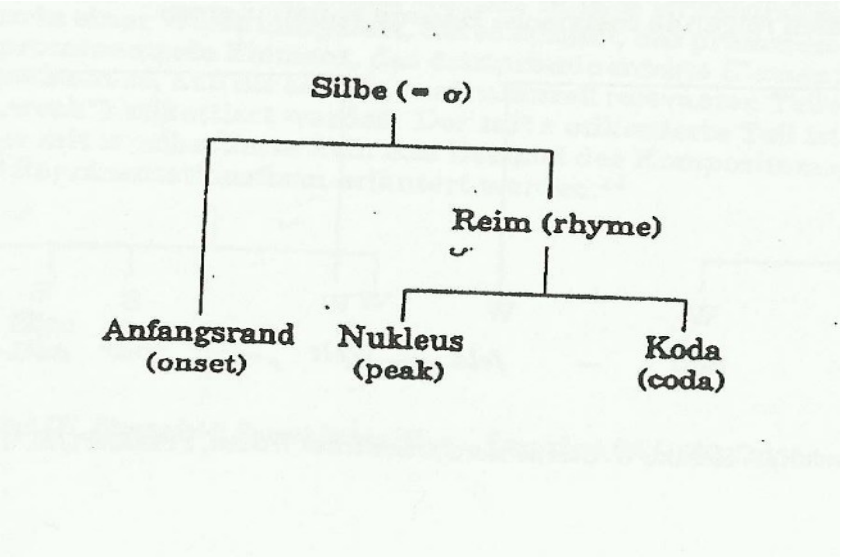
Prema strukturi na slici 3 je vidljivo da ulogu naglašenijeg sloga ima onaj kod kojeg se granaju strukture, odnosno kada se dijeli riječ unutar višesložne riječi na pojedine slogove. Time su riječi *„Fuß“* i *„spar“* naglašeniji slogovi od ostalih slogova unutar višesložne riječi. No, to pravilo ne vrijedi uvijek, što je vidljivo na primjerima

Altweibersommer i *Volkshochschule,* koji unatoč grananju imaju naglašeniji sloga na suprotnoj komponenti riječi.



Slika 3: Prikaz pravila izgovora u njemačkom jeziku [Gojmerac, 1992]

Kod riječi koje su sastavljene od dva sloga, moguće je naći naglasak na početnom ili krajnjem slogu. Zato naglasak ovisi o strukturi samih slogova. Razlikujemo između *teških* i *laganih* slogova. U *teške* slogove spadaju dugi vokal, dvoglas ili kratki vokal koji su zatim spojeni s jednim konsonantom. *Lagani* slogovi se označavaju kao otvoreni slogovi, koji poput jezgre ukazuju na kratki vokal. Kako bi se razumijela struktura slogova, potrebno je znati u kakve segmente je rastavljen slog[]. Na početnoj razini se slog sastoji od pristupa (njem. *Anfangsrand*) i rime (njem. *Reim*), dok se rima dalje grana na jezgru (njem. *Nukleus*) i odstup (njem. *Koda*) kao što je prikazano na slici 4 [Gojmerac, 1992: 86].



Slika 4: Univerzalna shema sloga u autosegmentnoj fonologiji[Gojmerac, 1992]

U njemačkom jeziku, kad je riječ o riječima sastavljenim od dva sloga, teški slog najčešće sadrži glavni naglasak. Kao primjer je moguće navesti par takvih riječi: *Protest* (Pro-test), *Papier* (Pa-pier), *Konsul* (Kon-sul), *Klima* (Kli-ma). Ovo pravilo ne

vrijedi univerzalno, jer postoje i razne riječi kojima je naglasak na krajnjem slogu, unatoč tome što su oba sloga u riječi *teška* (npr. *Kultur, Person*). Također postoje određene riječi sastavljene od tri sloga kojima je upravo središnji slog glavni (npr. *Veranda*). Ova kratka analiza ukazuje na kompleksnost njemačkog jezika i kako nije moguće ustanoviti podjelu naglaska u riječi samo na temelju par jednostavnih pravila. Takve probleme lingvisti objašnjavaju u opsežnim stručnim knjigama.

3.4 PRAVILA RASTAVLJANJA RIJEČI NA SLOGOVE

Prethodne cjeline su pokazale kako je rastavljanje riječi na slogove intuitivno jasan proces, ali je njegovo eksplicitno objašnjenje komplicirano i često puta nejasno zbog različitih elemenata koji se uzimaju u obzir. Do sada smo objasnili fonetske i fonološke pojedinosti kad je riječ o slogovima, dok ćemo se sada posvetiti definiciji pravila za rastavljanja riječi na slogove u njemačkom jeziku. Pojedina pravila vrijede za rastavljanje jednosložnih riječi, a druga za višesložne riječi. Prvo ćemo se osvrnuti na pravila za jednostavne riječi [Duden, 2009].

Pravilo 164:

1. Riječi sastavljene od više slogova se dijele pri laganom čitanju. Pojedina slova na početku ili kraju riječi se ne dijele. To vrijedi i za kompoziciju.
   - *Freun-de*, *Män-ner, Mül-ler, Bal-kon, aber-mals, Olym-pia-dorf*
2. Jedan jedini konsonant u unutrašnjosti riječi dolazi u pravilu u novu liniju; među više konsonanata biramo onoga koji je na samom kraju.
   - *tre-ten, nä-hen, Ru-der, An-ker, kämp-fen, schöns-te*
3. U slučaju da između dva vokala stoji kombinacija „ss“ umjesto „ß“, onda se riječ rastavlja između „ss“.
   - *Grüs-se* (umjesto: *Grü-ße*), *heis-sen* (umjesto: *hei-ßen*)

Pravilo 165:

1. Spojeni konsonanti *ch, ck i sch* (u stranim riječima: *ph, rh, sh i th*) se ne rastavljaju
   - *Bü-cher, Zu-cker, Ma-che-te, Pro-phet*
2. Dvoglasi *ai, au, äu, ei, eu, oi* se ne rastavljaju. Vrijedi i za riječi koje potječu iz Francuske
   - *Kai-ser, Trau-ung, Räu-ber, Cloi-son-né*
3. Nijema slova *e* i *i* također se ne rastavljaju. Nijemo *w* se tretira kao svaki drugi

konsonant.

- *Wie-se. Coes-feld, Trois-dorf*

Pravilo 166:

Strane riječi mogu imati određene grupe konsonanata nerastavljene

1. *bl, cl, fl, gl, kl, phl, pl* (primjer: *Pu-bli-kum* ili *Pub-li-kum*)
2. *br, cr, dr, fr, gr, kr, phr, pr, thr, tr, vr* (primjer: *Fe-bru-ar* ili *Feb-ru-ar*)
3. *gn, kn* (primjer: *Ma-gnet* ili *Mag-net*)

Pravila rastavljanja složenih riječi:

Pravilo 167:

1. Višesložne riječi i riječi s prednjim slogom se rastavljaju prema njihovim sastavnicama:
   - *Diens-tag, Stadt-staat, Neu-stadt, inns-bruck*
2. Pojedine sastavnice se dijele gore prikazanima pravilima
   - *Klei-der-schrank, Ho-sen-trä-ger, ge-ra-ten*
3. U slučaju da riječ više nije složena ili se ne smatra složenom, onda se rastavlja prema izgovoru slogova
   - *wa-rum* ili *war-um*, *ei-nan-der* ili *ein-an-der*, *He-li-kop-ter* ili *He-li-ko-pter*

Pravilo 168:

izbjegavaju se rastavljanja kod kojih se remeti slijed čitanja ili smisao riječi:

- *Spar-gelder* (umjesto: *Sparge-der)*, *be-inhalten* (umjesto: *bein-halten)*

4. IMPLEMENTACIJA PRAVILA RASTAVLJANJA U PROGRAMSKOME KODU

U ovom poglavlju se opisuju postupci rastavljanja riječi na slogove koji su implementirani u programu za automatsko rastavljanje. Preuzete su samo određene iznimke, no ne i sve, zbog njihovog velikog broja u njemačkom jeziku. iz svih do sada navedenih pravila i teorija rastavljanja slogova, odabrane su one koje se najčešće koriste pri automatskom postupku rastavljanja riječi na slogove. U to spada autosegmentalna fonologija, koja predstavlja ključnu CV-razinu između slogova i riječi. Također su uvedene mjere za rastavljanje jednosložnih i višesložnih riječi. Korištena su četiri osnovna pravila za rastavljanje riječi na CV-razinu:

1. Kombinacija „VKKV“ se rastavlja na „VK | KV“,
2. Kombinacija „KVKV“ se rastavlja na „KV | KV“,
3. Kombinacija „KVKK“ se rastavlja na „KVK | K“,
4. Kombinacija „KKVK“ se rastavlja na „KKV | K“.

Prvo pravilo osigurava da se rastavljanje vrši između dva konsonanta dok se vokali nalaze ispred i nakon konsonanata. Drugo pravilo vrši rastavljanje nakon vokala te prije konsonanta i time stvara simetričnu podjelu. Kod trećeg pravila se rastavljanje vrši tek na trećem mjestu nakon drugog pojavljivanje konsonanta. Četvrto pravilo također vrši rastavljanje na trećem mjestu, ali tek nakon pojavljivanja vokala. Za preciznije rastavljanje je uvedena lista skoro svih postojećih prefiksa i sufiksa. Oni ponajprije služe za rastavljanje višesložnih riječi na pojedine dijelove. Program pri dodjeljenju riječi iz rječnika prvo odvoja sve prefikse i sufikse koji čine tu riječ. Zadanu riječ zatim prikazuje kao zapis vokala i konsonanata. Dvoglasnike (odnosno diftonge) prikazuje skupno kao jedan zasebni vokal. Posebne kombinacije konsonanata također prikazuje skupno kao jedan konsonant, ali dodaje znakove „_“ na njihova mjesta kako bi ostala prepoznatljiva duljina riječi. U idućem koraku rastavlja strukturu vokala i konsonanata na 3 razine dok ne dođe do najjednostavnije. Nakon tog procesa ispisuje cijelu riječ rastavljenu na pojedine slogove i prebacuje se na iduću riječ. Riječi koje se sastoje od 3 ili manje riječi su izostavljene iz procesa rastavljanja riječi na slogove pošto već jesu slogovi. Ova korištena pravila su implementirane u program preko programskog jezika Python. Za detaljniji uvid u funkcioniranje cjelokupnog programa, postavljen je izvorni kod u dodatku ovog istraživanja.

5. REZULTATI

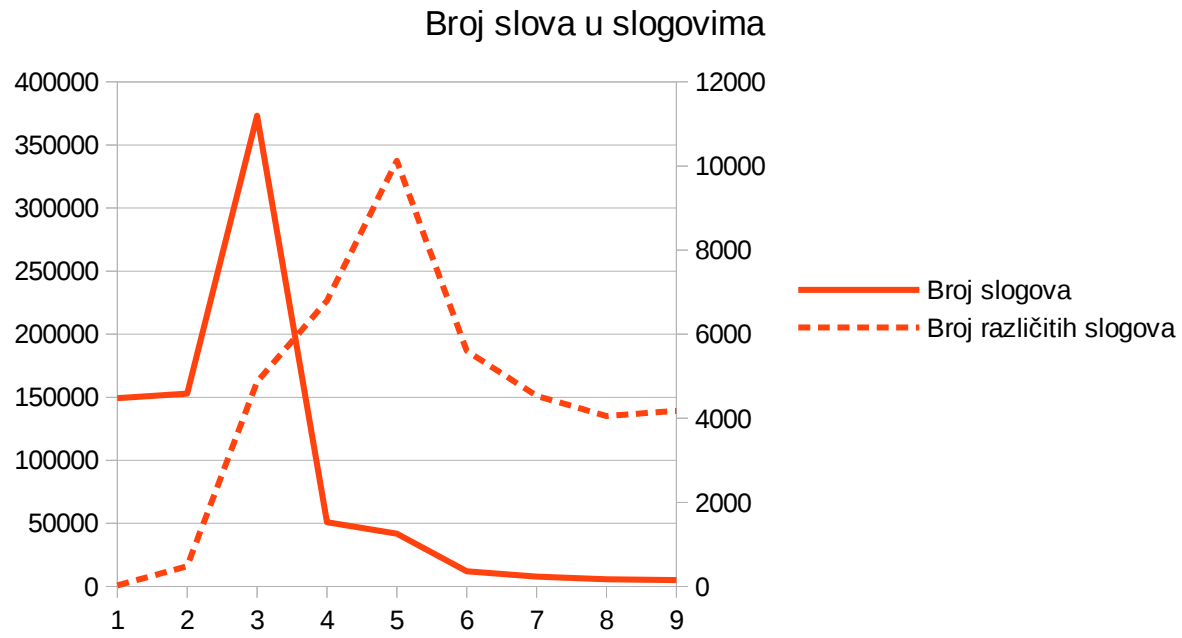
U ovom poglavlju prikazani su svi dobiveni rezultati tijekom istraživanja automatskog rastavljanja riječi na slogove. Svi rezultati su prikazani kroz statističku analizu u tablicama i grafu. Za provođenje statističke analize korišten je korpus riječi iz njemačkog rječnika [Padó, 2015], koji sadrži 208 977 leksema, što nije potpun broj svih leksema njemačkog jezika, ali predstavlja njihov reprezentativni udio. U ovaj korpus su uključeni i konvencionalni uzvici kao i riječi preuzete iz drugih jezika te vlastita imena. Provedena je analiza na temelju učestalosti određenih duljina slogova, analiza prema mogućem broju slogova unutar riječi, te zastupljenost određenih slogovnih modela unutar svake riječi.

5.1 ANALIZA REZULTATA AUTOMATSKOG RASTAVLJANJA

Prvi graf na slici 5 prikazuje distribuciju slogova unutar svake pojedine riječi. Iz zadanog korpusa postupak je rastavio riječi na 749,352 sloga, od kojih je pronađeno 40,660 različitih slogova. U postupku rastavljanja su izostavljene riječi koje sadrže manje od dva sloga jer su to ustvari samostalne i jednosložne riječi, te su već na razini slogova. U provjeri nije pronađena niti jedna riječ koja sadrži više od osam slogova. Na grafu je moguće vidjeti postepeni rast broja slogova sve do riječi koje sadrže 5 slova, što ujedno predstavlja najučestaliju skupinu riječi. Uočava se veliki broj riječi koje sadrže po četiri, pet i šest slogova. Među riječima koje sadrže najmanji broj slogova spadaju riječi s dva i osam sloga.

Slika 5: Graf broja slogova u riječi

Iduća analiza je provedena u cilju pronalaska zastupljenosti slogova prema njihovoj duljini, odnosno od koliko se slova sastoje. U analizi su korišteni podaci ukupnog broj slogova i različitih slogova. Rezultati, koji su prikazani na slici 6, ukazuju na poprilično velik broj slogova koji se sastoje od 3 slova (njihov postotak iznosi oko 46%). Dok su drugi najučestaliji slogovi oni koji se sastoje od 2 slova (njihov postotak iznosi oko 19%). Uočljiv je nagli pad frekvencije slogova nakon onih koji se sastoje od 3 slova. Već slogovi koji se sastoje od 4 slova zauzimaju svega 6%, dok su slogovi koji se sastoje od 9 slova najrjeđi. Promatrajući broj različitih slogova uviđaju se suprotnosti u frekvenciji slogova. Ukupni broj slogova dostiže svoj najveći postotak frekvencije u slogovima koji su sastavljeni od 3 slova, dok različitih slogova od 3 slova ima svega 11%. Najveći postotak različitih slogova čine slogovi sastavljeni od 5 slova (njihov postotak iznosi skoro 25%), kojih u ukupnom razmatranju ima oko 5%. Ovaj dio analize može ukazivati na podudaranje, odnosno preklapanje, sličnih pravila rastavljanja slogova (korištena pri implementaciji u programa), koja ponekad moraju biti u pravilnom redoslijedu kako bi rastavljanje slogova bilo ispravno.



Slika 6: Graf ukupne frekvencije svih i različitih slogova

Zadnja analiza je napravljena za provjeru zastupljenosti svakog pojedinog modela sloga u zadanom korpusu. Za ovu analizu su iskorištene tri razine rastavljanja za svaku riječ, kako bi se obuhvatio cjelokupni skup modela sloga. Postoji veliki postotak slogova s zadanim modelom „CVC“ (44,8%). To je shvatljivo jer je na prijašnjem grafu ustanovljeno da je najveći postotak slogova sastavljen od samo 3

slova. U ovu statistiku ulaze i slogovi koji se sastoje od 4 slova, pošto vokal može predstavljati dvoglas. Modeli sloga „CV“ i „CCV“ odmah prethode modelu „CVC“ (oko 18% i 10%). Zbrojem preostalih 11 modela sloga dobio bi se postotak od oko 20%, što ukazuje na najveću zastupljenost 3 modela sloga. Najmanje zastupljeni model sloga je „VCC“ koji se svega 14 puta uspio realizirati, kao što je i prikazano u tablici 2. U tablici je prikazan udio za svakog od 16 reliziranih modela slogova.

Tabela 2: Frekvencija modela sloga

Model sloga	**Korpus: Rječnik**	
	Broj slogova	%
CVC	496564	44.8
CV	209764	18.95
CCV	112423	10.16
CCVC	74914	6.77
CVCC	60792	5.49
VC	45039	4.07
CVCCC	38657	3.49
CCCV	31429	2.84
CCCVC	22703	2.05
CCCCV	11937	1.07
V	851	0.07
CCVCCC	667	0.06
CCVCC	470	0.04
VCCC	166	0.01
CCCVCC	74	0.006
VCC	14	0.001

5.2 PROVJERA I PROCJENA POGREŠKE

U ovom dijelu rada se opisuju postupci ručne provjere automatskog rastavljanja riječi na slogove i procjena pogrešaka koje se pojavljuju. Iz korpusa svih dobivenih riječi slučaknim odabirom je izdvojeno 100 riječi za provođenje evaluacije automatskog rastavljanja. Provjera točnosti automatskog rastavljanja je provedena uz pomoć pravila rastavljanja na slogove, koja se nalaze unutar rječnika njemačkog jezika [??]. Kod riječi koje se sastoje do maksimalno 3 sloga, uključujući prefikse i sufikse, ustanovljena je apsolutna točnost pri svakom pokušaju. Najveći problem

predstavljaju višesložne riječi. Točnost programa na temelju ovih izabranih 100 riječi je iznosila točno 60%. Razlog tome je podudaranje sličnih zadanih pravila kojima program nije mogao odrediti prioritete. Moguće iznimke pravila kod pojedinačnih riječi se mogu uzeti u obzir. Rastavljanje jednostavnih riječi i riječi stranog podrijetla ne predstavljaju nikakve probleme programu. Problem pri rastavljanju višesložnih riječi nastaje kad riječi sadrže prefikse i sufikse unutar svoje strukture. U budućem radu bit će se potrebno posvetiti definiranju i implementaciji dodatnih pravila koji će otkloniti ovu pogrešku.

6. ZAKLJUČAK

Rastavljanje riječi na slogove se može činiti poput intuitivno jasnog ili trivijalanog postupka. Međutim problemu rastavljanja su se posvećivali mnogobrojni lingvisti tijekom proteklih stoljeća. To je i dalje bitan dio našeg jezika jer boljim razumijevanjem vlastitog jezika lakše razumijemo sebe. Ovakvi osnovni elementi jezika su od važnosti u daljnjem razvoju tehnologije i samog jezika. U ovom radu je objašnjen postupak rastavljanja riječi na slogove preko implementacije u programskom kodu u Pythonu. Za razumijevanje pravila slogova, potrebno je znati znanstvene discipline fonetike i fonologije, koje su doprinjele današnjem saznanju tih osnovnih elemenata jezika.

Uz pomoć programa za automatsko rastavljanje riječi na slogove bila je omogućena statistička analiza pojedinih detalja slogova unutar njemačkog jezika. Uspješno je analizirana učestalost broja slogova unutar svake riječi, duljina svakog pojedinog sloga i frekvencija pojavljivanja 16 modela sloga u svakoj riječi. Na kraju analize rezultata je ustanovljeno:

- da su slogovi duljine tri slova najučestaliji (46%), te da slogovi većih duljina nisu zastupljeni u tolikoj mjeri (ukupno oko 20%);
- da ima najviše riječi koje sadrže po četiri i pet različitih slogova (oko 40%);
- da je najveći broj različitih slogova među slogovima sastavljenih od 5 riječi (oko 25%)
- da je „CVC“ najzastupljeniji model sloga (44%) što odmah potvrđuje rezultate o slogovima duljine tri slova.

Iako se slogovi smatraju osnovnim elementima svake riječi, oni istovremeno iskazuju strahovitu posvećenost detaljima u kojima se odražava njihova kompleksnost. Unatoč brojnim pravilima rastavljanja riječi na slogove u njemačkom jeziku, i dalje postoji puno iznimki koje nisu jednostavne za implementirati unutar automatskog rastavljanja.

LITERATURA

Gojmerac, M. (1992), *Einführung in die Linguistik*, Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu, Zagreb.

Hansen, Gallmann, Eisenberg, Fiehler, Peters, Nübling, Barz, Fritz. (2009), *Duden 04. Die Grammatik*, Dudenredaktion, Zürich.

Duden (2011), *Deutsches Universalwörterbuch*, Dudenredaktion, Zürich.

Bird S, Klein E, Loper E. (2009), *Natural Language Processing with Python*, O'Reilly Media.

Tapken, H. (2013), *Natural Language Processing Sentimentanalyse*, http://www.ecs.hs- osnabrueck.de/44188.html (posjet 09/2015).

Edelman, S. Waterfall, H. (2007), *Behavioral and computational aspects of language and its acquisition*, ScienceDirect.

Liang Mark, F. (1983), *Word Hy-phen-a-tion by Comput-er*, Stanford University.

Padó, S. (2015), *DerivBase*, http://www.ims.uni-stuttgart.de/permalink/56cc6c89-c421-11e4-a5e6-000e0c3db68b.html (posjet 09/2015).

DODATAK 1

```
#!/usr/bin/python
# -*- coding: latin-1 -*-
import os, sys
import string
import re

def slogovi(rijec):
    samoglasnici = ['a', 'e', 'i', 'o', 'u']
    comb = ['sch', 'ch', 'lich', 'ich', 'tich', 'ss', 'ft']
    pravila = {'VKKV' : 'VK|KV',
               'KVKV' : 'KV|KV',
               'KVKK' : 'KVK|K',
               'KKVK' : 'KKV|K'
               }
    temp = ''

    rijec = rijec.lower()

    if len(rijec) <= 3:
        return rijec

    prefiksi = ['ab', 'au', 'an', 'auf', 'ma','re', 'nach','fest', 'hin', 'be', 'fehl',
 'er', 'ein', 'durch', 'um', 'zer', 'zu', 'vor', 'ver', 'ge', 'aus', 'wahr', 'zy', 'te']
    nova = ''
    for p in prefiksi:
        if rijec.startswith(p):
                nova = rijec[:len(p)] + '-' + rijec[len(p):]

    sufiks = ['lich', 'lein', 'isch', 'sch', 'sche', 'bar', 'tich', 'fon', 'sam', 'chte',
 'nist', 'ne', 'de', 'end', 'licht', 'bold', 'chen', 'de', 'e', 'er', 'el', 'ei', 'el',
 'erl', 'la', 'le', 'tje', 'je', 'ken', 'ig', 'ke', 'heit', 'keit', 'icht', 'ian', 'jan',
 'i', 'in', 'lein', 'ler', 'ling', 'ner', 'nis', 'rich', 'sal', 'schaft', 'sel', 'tel',
 'tum', 'ung', 'ing']
    for p in sufiks:
        if len(nova) < 1: #pogleadati za empty()
                nova = rijec
        if nova.endswith(p):
                nova = nova[:-len(p)] + '-' + nova[-len(p):]
                print nova
```

```
35
36      r = nova.split('-')
37      if len(r) == 3:
38          rijec = r[1]
39      else:
40          rijec = r[0]
41
42      for x in comb:
43          nova = re.sub(x, len(x)*'_', rijec)
44
45      for char in nova:
46          if char in samoglasnici:
47              temp += 'V'
48          elif char == '_':
49              temp += '_'
50          else:
51              temp += 'K'
52
53      pomocni = re.sub('VV', 'V', temp)
54      print temp, pomocni
55
56      razdjela = ''
57      for key, value in pravila.iteritems():
58          if key in pomocni:
59                  razdjela = re.sub(key, value, pomocni)
60                  pomocni = razdjela
61                  print razdjela
62
63      stop = False
64
65      j = 0
66      for i, word in enumerate(razdjela): #KVK__
67          try:
68              j = i
69              if rijec[i] in samoglasnici and rijec[i+1] in samoglasnici:
70                  j = i+2
71                  if word == '|':
72                          rijec = rijec[:j+1] + '-' + rijec[j+1:]
73              else:
74                  if word == '|':
75                          rijec = rijec[:j] + '-' + rijec[j:]
76
77          except IndexError:
78              break
79
80      if len(r) == 3:
81          r[1] = rijec
82      else:
83          r[0] = rijec
84      print '-'.join(r)
85
86  f = open('/home/johnny/Dropbox/Završni Rad/Novo/noviRez.txt', 'r')
87  for line in f:
88     slogovi(line)
89  f.close()
```