Razvoj alata za proveru i korekciju teksta zasnovan na grafovima i stablima pretraživanja

Jovana Kitanović, Dražen Drašković, Maja Vukasović, Sanja Delčev

***Apstrakt* – Istraživanje se bavi razvojem alata za proveru teksta tokom samog unosa, i korekciju teksta, prema pravopisu nemačkog jezika. Istraživanju je prethodila analiza sličnih alata kao što su Google Translate, Grammarly, Insta Text, koji nisu zadovoljili sve postavljene kriterijume. Prvi izazov bio je formiranje velikog skupa podataka, odnosno rečnika nemačkih reči, iz dostupnih izvora. Formiran je rečnik sa 357 hiljada reči. Drugi izazov bio je odabir struktura podataka koje će se koristiti u alatu i analiza mogućih tehnika učitavanja reči iz baze u odabranu strukturu podataka, kako bi učitavanje baze bilo što efikasnije. Alat je razvijen kao veb platforma, koja radi pouzdano i postiže najbolje moguće performanse.**

***Indeks termina* – grafovi, stabla pretraživanja, baze podataka, rečnik, nemački jezik.**

I. Uvod

U razvoju alata za proveravanje unetog teksta na nekom jeziku veoma je bitna efikasnost i izbor struktura podataka je od velike važnosti. Neki jezici su prilično komplikovani u gramatici i imaju veliki broj reči, pa je izazov praviti softverske alate za takve prirodne jezike [1].

Nemački jezik ima četiri padeža: nominativ, genitiv, dativ i akuzativ, a slično kao i u srpskom jeziku, reči koje trpe padežne promene su imenice i član uz imenicu, zamenice, pridevi i brojevi. Pored imenskih reči, gramatičku promenu trpe još glagol i prilozi. Nemački jezik ima raznovrsan skup glagolskih vremena i rečeničnih konstrukcija od kojih zavisi i oblik glagola koji će se u takvim rečenicama koristiti [2]. U tabeli 1, dat je ilustrativni primer raznovrsnosti glagolskih oblika i nekih vremenskih konstrukcija za glagol *kochen* – kuvati. Da bi provera pravopisa mogla da bude implementirana, potrebno je napraviti skup poznatih reči u odnosu na koje će se uneti tekst porediti. Postavljeni su sledeći kriterijumi: 1) da svaka imenica ima rod, nastavak za genitiv i nastavak za množinu; 2) da za svaki glagol postoji podatak da li je pravilan, da li je povratni, koji je oblik glagola u trećem licu jednine, kao i njegov pomoćni glagol; 3) sve ostale reči da imaju oznaku vrste kojoj pripadaju.

Kako već gotovi skupovi podataka nisu zadovoljavali postavljene kriterijume, bilo je potrebno pronaći pouzdan izvor iz koga se reči mogu preuzeti uz formiranje velikog skupa reči. Odabrane je nemačko-nemački rečnik „*Duden Deutsches Universalwörterbuch*“ [3].

U drugoj sekciji opisan je proces prikupljanja podataka. Treća sekcija daje opis izazova kako odabrati najpogodniju strukturu podataka. U sekciji četiri opisan je proces učitavanja rečnika. Peta sekcija opisuje korisnički interfejs realizovanog alata za proveru i korekciju teksta unetog na nemačkom jeziku. Na kraju rada dat je zaključak.

TABELA I
PRIMER RAZNOVRSNOSTI GLAGOLSKIH OBLIKA

Zamenica	Prezent	Preterit	Futur I
ich	koche	kochte	werde kochen
du	kochst	kochtest	wirst kochen
er/sie/es	kocht	kochte	wird kochen
wir	kochen	kochten	werden kochen
ihr	kocht	kochtet	werdet kochen
Sie/sie	kochen	kochten	werden kochen

II. Proces prikupljanja podataka

Proces generisanja skupa reči, odnosno rečnika odvijao se u nekoliko faza: učitavanje rečnika u PDF formatu i čuvanje reči iz njega u odvojenim tekstualnim fajlovima, izolovanje reči iz takvog rečnika, pronalaženje gramatike za izolovane reči, filtriranje reči po vrsti, prečišćavanje podataka, dodavanje padežnih oblika imenskih reči i pronalaženje svih glagolskih oblika izolovanih glagola.

A. *Pronalaženje reči i dodela gramatike rečima*

Odabrani rečnik ima 2132 stranica, pa bi proces pretraživanja bio dug. Zato je iz rečnika za svako početno slovo reči nemačkog jezika, formiran po jedan fajl, što je rezultovalo sa 25 manjih pretraživih rečnika (reči na slova X, Y i Z su smeštena u jedan fajl; takođe u istom fajlu su reči koje počinju sa a i ä, ili o i ö). Naredni korak podrazumevao je izolovanje pojedinih reči. Odabrani rečnik je formatiran tako da se reči prema formatu mogu izolovati. Reči su takođe podeljene na slogove. Čitajući reč i nekoliko narednih reči, koje daju objašnjenja, odabranoj reči se dodeljuje gramatika. Na primer, za reč *Reiserverbot*, čitamo reč i dodeljujemo rod:

Rei|se|ver|bot, das

U slučaju reči *Phänomen*, gramatika se dodeljuje čitanjem tri naredne reči njenog objašnjenja, koje označavaju rod, nastavak za oblik u genitivu i nastavak za oblik reči u množini:

Phä|no|men, das; -s, -e

Kako prilikom pronalaska reči ne možemo znati koliko narednih reči je potrebno pročitati, to je uočena dalja pravilnost da se osnovna objašnjenja svih reči završavaju simbolom dvotačka. Tako će se za svaku reč inicijalno pročitati više teksta nego što je potrebno, dok će se u narednim ciklusima filtriranja podataka reči ona svesti na oblik objašnjen u prethodnom pasusu.

Problem se javio prilikom čitanja reči koje ne mogu biti podeljene na slogove, odnosno prilikom čitanja jednosložnih

Jovana Kitanović, Dražen Drašković, Maja Vukasović and Sanja Delčev are with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11020 Belgrade, Serbia (e-mail correspondence: drazen.draskovic@etf.bg.ac.rs), ORCID ID (https://orcid.org/0000-0003-2564-4526, 0009-0008-8957-2254)

 ISBN 978-86-7466-969-3

reči. Naime, takve reči ne mogu biti nedvosmisleno locirane u tekstualnim fajlovima, jer se potencijalno koriste za objašnjavanje drugih reči rečnika, stoga su one smeštene u posebne tekstualne fajlove i obrađene drugačijim pristupom. Usvojeno je da su jednosložne reči u rečniku upisane u formatu: reč, član; -nastavak za genitiv, -nastavak za množinu.

B. *Obrada složenica*

Nemački jezik je bogat složenicama, od kojih se veliki broj nalazi u rečniku nemačkog jezika. Međutim, u odabranom rečniku se one ne nalaze u potrebnoj formi, tačnije, osim roda, gramatički nastavci nisu zapisani, te je za takve reči potrebno pronaći njihov koren, kako bi im se nedostajući gramatički nastavci dodali. Složenice se u rečenicama menjaju kroz gramatiku po pravilima koje prati koren reči. Na primer: *Autobahn* – autoput, je reč sastavljena od dve reči, *Auto* - automobil i *Bahn* - traka, put:

Au|to|bahn, die: Schnellstrase, die kreuzungsfrei
u. zwei- od. mehrspurig nur fur bestimmte Kraftfahrzeuge zugelassen ist.
Au|to, das; -s, -s [Kurzf. von ↑Automobil]
Bahn, die; -, -en

Ovakvi slučajevi, koji čine većinu imenica nemačkog jezika, se obrađuju pronalaskom najvećeg preklapanja poslednje reči. Na primeru reči *Softwareentwickler* - osoba koja razvija softver, u reči možemo primetiti reč *er* značenja on, međutim najduža reč je *Entwickler* - osoba koja se bavi razvojem, stoga se ona uzima kao koren i po njenoj gramatici se posmatra i složenica:

Soft|ware|ent|wick|ler, der: jmd., der im Bereich
der Softwareentwicklung arbeitet.
Ent|wick|ler, der; -s, -
Er, der; -, -[s]

C. *Obrada glagola*

Pronalaženje glagola u rečniku se vrši prema istom principu pretraživanja višesložnih i jednosložnih reči u tekstualnim datotekama. Format u kom su glagoli u rečniku upisani je nešto drugačiji od ostalih vrsta reči. Naime, za svaki glagol je važno znati da li je glagol pravilan ili nepravilan, a ukoliko je pravilan, da li je jak ili slab, koji pomoćni glagol koristi i da li je povratni. Sve informacije su zapisane u rečniku iza glagola u formatu:

glagol ⟨st./sw./unr. V. ; hat/ist ⟩

ili ukoliko je glagol povratni:

glagol, sich ⟨st./sw./unr. V. ; hat/ist⟩

Poznavanje informacija o glagolima je od velikog značaja za generisanje svih glagolskih oblika u kojima se glagoli mogu naći. Prilikom filtriranja glagola, doneta je odluka da se oni odvoje prema tome da li su pravilni ili ne, a zatim pravilni da se razdvoje na jake i slabe. Obe kategorije glagola, jaki i slabi, kako pripadaju grupi pravilnih glagola imaju tačno definisana pravila po kojima se, kroz glagolske oblike menjaju. Međutim, nepravilni glagoli ne podležu pravilima, te promena kroz glagolska vremena ne može biti automatizovana. Stoga je odlučeno da se glagolski oblici glagola preuzimaju sa internet stranice „*Pons*“ internet rečnika koristeći veb indekser i veb parser [4-5].

Ukupan broj glagola dobijen ovim metodama prelazi broj od 131 hiljade, međutim samo je 65 hiljada jedinstvenih oblika glagola u rečniku. Objašnjenje za veliku razliku u broju glagola i broju glagola upisanih u rečnik su glagoli sa razdvojivim prefiksima, kojih u rečniku ima oko 50% od ukupnog broja glagola u rečniku. Nakon završene obrade, ukupan broj reči preuzet iz rečnika je 357 hiljada, od čega 193 hiljade imenica, oko 65 hiljada glagola, oko 7 hiljada priloga, više od 88 hiljada prideva, dok preostali skup čine članovi, zamenice, imena, skraćenice, predlozi, brojevi i veznici.

III. ANALIZA PROBLEMA ODABIRA STRUKTURA PODATAKA

Većina alata koja je analizirana u ovoj oblasti, poput Grammarly, Instatext, Google Translate, radi proveru reči, tek nakon nekog vremenskog intervala, a ne nakon unosa svakog pojedinačnog karaktera. Takođe, kod prva dva alata neophodno je kucanje malo većeg teksta radi provere ispravnosti [6,7]. Od analiziranih alata, jedino je Google Translate podržao nemački jezik.

Prilikom dizajniranja ovog alata, zamisao je bila da se omogući provera teksta tokom samog unosa, odnosno nakon svakog unetog ili obrisanog karaktera. Prvi mogući pristup je kreiranje upita nad bazom, za svaku proveru, a drugi pristup je kreiranje strukture koja će tokom rada alata čuvati sve reči baze.

Prvi pristup bi morao imati izuzetno brzu komunikaciju sa bazom, odnosno veoma brzo kreiranje upita, slanje upita, obradu od strane baze, prihvatanje odgovora od baze i njegovu obradu. Broj takvih upita bi bio veliki, i varirao bi od korisnika do korisnika, u odnosu na brzinu kucanja. Potencijalni problemi koji bi se javili bili bi preoptećerenje i nedovoljna brzina obrade zahteva.

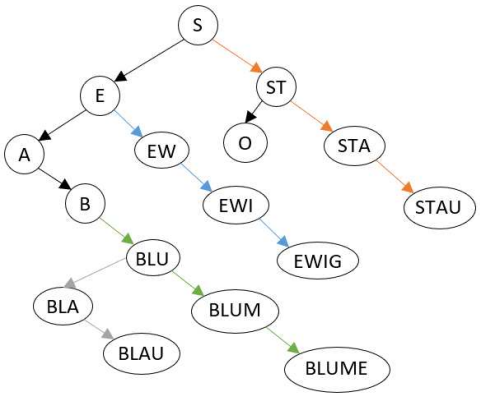
Drugi pristup bi kreirao strukturu podataka koja bi skladištila sve reči tokom rada aplikacije, a zatim po potrebi vršila pretragu strukture na reči ili delove reči i dohvatale bi se vrsta ili vrste reči, ukoliko ih ima više. Strukture koje bi zadovoljile potrebe ovog alata su grafovi, stabla opšteg pretraživanja, heš tabele ili modifikovana stabla binarnog pretraživanja. Dodatni zahtev koji struktura treba da zadovolji u svrhu brze pretrage je da detektuje neispravne reči što ranije, da ima mogućnost upisa nove reči nakon formiranja strukture i da omogućava čuvanje dodatnih informacija, u slučaju rečnika je to vrsta reči.

Zaključeno je da korišćenje heš funkcija može biti previše komplikovano za brzo pretraživanje. Samo pronalaženje, a kasnije izračunavanje heš funkcije može biti zahtevno, stoga i obrada zahteva za proverom reči može biti usporena. Dodatno, ukoliko bismo koristili heš funkcije za pretragu rečnika i ukoliko bi se kolizije dešavale (mapiranje rezultata heš funkcija različitih reči u isti ulaz tabele), bilo bi potrebno dodatno preračunavanje kako bi se nova reč u strukturu ubacila na slobodno mesto ili kako bi se postojeća reč pronašla.

Binarna stabla su pogodna za pretragu, međutim, kako svaki čvor može imati svega dva podstabla i to takva da su ključevi levog podstabla manji od korena, a svi ključevi desnog podstabla veći od korenog ključa, to takođe nije dobra struktura za ovu realizaciju. Dodatni uslov je da ključevi stabla moraju biti jedinstveni. Da bi se ovaj uslov ispunio, ključevi bi morali da budu kombinacija onoliko početnih slova reči kolika je dubina stabla, čime se dobijaju jako dugački ključevi u već previše razgranatoj strukturi. Da bi struktura bila optimizovana, tačnije, da bi se brzo odredila ispravnost čestih reči, potrebno ih je čuvati blizu korena stabla. Takav pristup zahteva da se tokom korišćenja alata struktura prilagođava i reorganizuje, a to može biti skup

 ISBN 978-86-7466-969-3

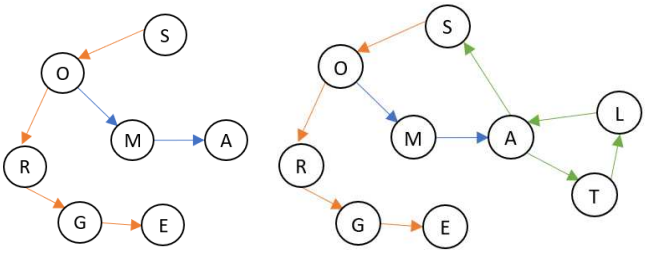
proces imajući u vidu razgranatost strukture i broj reči koje bi struktura morala da podrži. Prikaz opisa binarnog stabla sa slovom S u korenu i četiri reči, *ewig* - večno, *Stau* - gužva u saobraćaju, *blau* - plavo, *Blume* - cvet, dat je na slici 1.



Sl. 1. Izgled rečnika u strukturi binarnog stabla

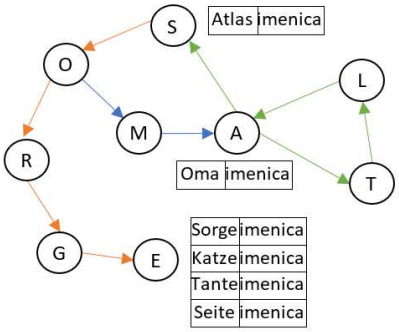
Nakon analize zahteva koje struktura treba da zadovolji i eliminacije nedovoljno pogodnih struktura, izbor je sveden na grafove i stabla opšteg pretraživanja.

Najbitniji zahtev koje struktura treba da zadovolji jeste brzina kojom može da odredi da li je reč ispravna ili ne, stoga reč možemo posmatrati kao putanju koju treba proći kroz čvorove grafa, a slova u reči kao čvorove. Ukoliko su svi čvorovi dostižni reč je ispravna, a ukoliko makar jedan čvor ne može biti dostignut reč je neispravna. Takvim pristupom ćemo veoma rano, u procesu provere reči eliminisati one koje nisu ispravne. Međutim, kako je graf struktura koja može imati cikluse, postoji mogućnost da čak i neispravne reči budu detektovane kao ispravne. Za potrebe objašnjenja problema detekcije neispravne reči kao ispravne, koristiće se nekoliko reči nemačkog jezika: *Sorge* – briga, *Oma* - baka, *Atlas* – atlas. Ako skup reči ima samo prve dve reči, graf će izgledati kao na slici 2 (levo), ali će se u tom slučaju ispravnom rečju smatrati i reč *soma*, jer postoji putanja S-O-M-A kroz čvorove grafa. Pomenuti problem detekcije neispravnih reči kao ispravnih se pogoršava povećanjem broja reči, a time dodavanjem grana između čvorova (slika 2, desno).



Sl. 2. Izgled grafa sa dve reči (levo) i nakon dodavanja reči Atlas (desno)

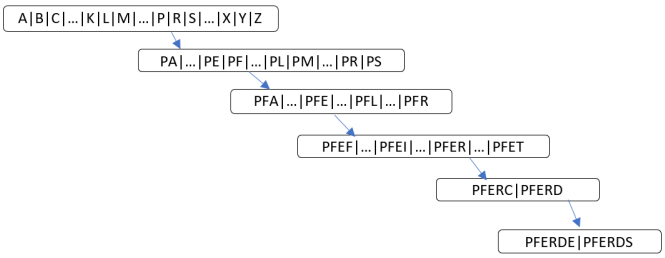
Potencijalno rešenje pomenutog problema, kao i zadovoljenje zahteva za čuvanjem dodatnih podataka u smislu vrste reči, bilo bi dodavanje liste ispravnih reči i njenih vrsta svim čvorovima koji mogu biti poslednja slova u reči, kao što je prikazano na slici 3. Uzimajući u obzir da reči nemačkog jezika mogu biti izuzetno duge i da će upisom svih reči rečnika u strukturu svaki čvor postati dostižan iz svakog drugog čvora, dolazimo do zaključka da struktura grafa samo može smanjiti skup reči koje pretražujemo kako bismo utvrdili ispravnost reči, a ne i sprečiti tu situaciju eliminacijom reči zbog nemogućnosti pronalaska putanje kroz graf, kao što je bila inicijalna zamisao.



Sl. 3. Izgled grafa sa kolekcijom reči

Stablo opšteg pretraživanja koje se za potrebe alata za proveru gramatike razmatra je m-arno stablo. M-arno stablo je stablo kod kojeg je stepen svakog čvora manji ili jednak stepenu m. Ključevi stabla su uređeni u rastućem poretku. Jedan ključ razdvaja dva podstabla, koja su takođe stabla m-arnog pretraživanja, tako da su svi ključevi podstabla levo od ključa manji od njega, dok su svi ključevi desnog podstabla veći od njega.

Kako bi se zadovoljila pravila m-arnog stabla, reči rečnika se postepeno razlažu i upisuju kao ključevi. Stepen stabla će biti broj slova nemačkog alfabeta, što iznosi trideset. Da bi se zadovoljio uslov da desno podstablo korena bude takođe m-arno stablo i da svi ključevi budu veći od korenog ključa, naredni nivo stabla će predstavljati sve moguće kombinacije prva dva slova nemačkih reči takvih da je prvo slovo koreni ključ. Treći nivo će sadržati ključeve koji su kombinacija prethodnog ključa, tačnije, prethodna dva slova i trećeg slova koje može da figuriše kao treće slovo reči u kombinaciji sa prethodna dva slova. Svaka reč rečnika će biti podeljena na pomenuti način sve dok cela ne bude uneta u stablo, što se može videti na primeru reči *Pferd* - konj, na slici 4.



Sl. 4. Izgled rečnika u formi m-arnog stabla

Ovako formirano stablo će omogućiti da se reči koje nisu ispravno unete odbace kao netačne, čim se prolaskom kroz stablo utvrdi da takva kombinacija slova u stablu ne postoji. Umetanje nove reči u stablo neće poremetiti postojeću strukturu stabla, dok dodatne informacije o reči mogu da se čuvaju uz sam ključ, te se može zaključiti da prikazana struktura u potpunosti zadovoljava postavljene zahteve.

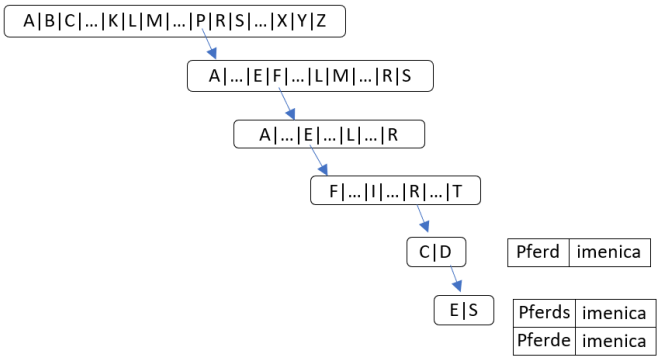
Nedostatak ovakvog pristupa je u samom ključu, jer povećanjem dužine reči, ključevi postaju glomazniji, a samim tim komplikovaniji za obradu. Na primer, ukoliko bi reč bila *Schauspieler* - glumac, prostirala bi se na dvanaest nivoa, od čega bi poslednji nivo imao ključ dug dvanaest slova. Pri tome, reč *Schauspiel* - gluma, može biti deo dužih složenica kao na primer *Schauspielkunst* - umetnost glume

 ISBN 978-86-7466-969-3

ili *Schauspielkarriere* - glumačka karijera kada bi ključevi imali petnaest odnosno osamnaest slova ili duže. Struktura podataka koja rešava problem glomaznih ključeva, a čuva sve potrebne pozitivne osobine m-arnog stabla jeste trie stablo digitalnog pretraživanja.

Trie (digitalno / radiks) stablo je organizovano tako da je ključ svakog nivoa zapravo jednak putanji od korena do tog ključa. Struktura stabla nije organizovana po principu većih i manjih vrednosti ključeva narednog nivoa u odnosu na ključ trenutnog nivoa, već samo polje ključa sadrži pokazivač na naredni nivo. Struktura trie stabla zadovoljava sve potrebe alata za proveru pravopisa, odnosno zadovoljava sve postavljene zahteve. Omogućava da neispravno napisane reči budu detektovane tokom samog prolaska kroz stablo, omogućava lako dodavanje novih reči u strukturu, a pored toga omogućava čuvanje dodatnih podataka uz ključ stabla.

Kako bi se prostor za strukturu ekonomično koristio, umesto fiksne veličine čvora što je osobina trie stabla, veličina čvora će varirati zavisno od potrebe. Tako na primer, ukoliko posle prvog slova reči nemačkog jezika može da sledi samo nekoliko slova alfabeta, veličina čvora neće biti trideset nego upravo onoliko koliko je dovoljno da se predstave sve reči. Primer strukture je prikazan na slici 5 za istu reč *Pferd* - konj, na kojoj je ilustrovana i struktura m-arnog stabla.



Sl. 5. Izgled rečnika u formi trie stabla

IV. Proces učitavanja rečnika

Mogući pristupi učitavanja rečnika su učitavanje čitavog rečnika odjednom ili učitavanje rečnika na određeno slovo, onda kada korisnik prvi put unese reč na to slovo. Jedno od alternativnih mogućnosti bilo bi i kombinovanje najboljih odlika ova dva pristupa.

A. *Postepeno učitavanje rečnika*

Postepeno učitavanje rečnika je zamišljeno tako da je inicijalno struktura rečnika prazna i ne počinje da se popunjava do trenutka kada korisnik počne sa unosom teksta. Tada se za svako početno slovo unete reči kreira nit koja vrši učitavanje reči na to slovo u rečnik. Problemi koji se prilikom ovakvog pristupa javljaju jesu nejednaka brzina učitavanja rečnika za sva slova.

Konkretno, broj reči u rečniku je najveći za rečnik na slovo S sa 42 hiljade reči, a najmanji za slova X sa 77 reči i Y sa 67 reči, te će i brzina učitavanja rečnika biti proporcionalna. Nijedna reč tokom tog procesa učitavanja, ukoliko rečnik za to slovo nije napravljen, neće biti proverena. Ovo je prihvatljiv slučaj iz razloga što korisnik neće čekati dugo da reči budu proverene, ali će osetiti nakratko zastoj u radu alata.

Veći problem jeste slučaj kada korisnik u polje za unos teksta unese odjednom čitav tekst: nit koja je pokrenuta da proveri gramatiku će detektovati da je rečnik prazan, te će za svako početno slovo koje uoči, pokrenuti nit za učitavanje strukture rečnika. Potencijalno će se pokrenuti 30 niti za svih 30 slova alfabeta, čime će se povećati ukupno vreme koje korisnik čeka na proveru gramatike, što se kosi sa osnovnim ciljem alata. Osim toga, korisnik ne treba da ima utisak da treba da čeka alat niti da postane nestrpljiv i izgubi poverenje u performanse, sposobnosti i komfor alata.

B. *Celovito učitavanje rečnika*

Celovito učitavanje pokreće kreiranje rečnika za svih 30 slova odjednom u paralelnoj *for* petlji. Ovakvim pristupom se postiže to da korisnik ima čitav rečnik spreman pre nego što unese prvu reč. Prednost ovakvog pristupa je što korisnik neće imati nikakvo kašnjenje prilikom provere reči niti će osetiti bilo kakvo neočekivano ponašanje alata prilikom unosa čitavog teksta odjednom. Nasuprot očiglednih prednosti celovitog učitavanja rečnika, nedostatak je to što će samo pokretanje alata potencijalno trajati nešto duže, što ponovo kod korisnika može izazvati nestrpljivost, ali bi se kroz stabilan alat potpuno spreman za rad odmah nakon učitavanja taj nedostatak ublažio.

U idealnim uslovima kada se ne bi moglo desiti da konekcija sa bazom prestane, ili za slučaj da nije u stanju da obradi filtriranje reči po početnom slovu, alat bi se pokretao konzistentnom brzinom prilikom svakog pokretanja. Ali, takvi slučajevi se mogu dogoditi i dovode do usporenog pokretanja aplikacije u nepredvidivim trenucima, te je najsigurniji pristup kombinacija celovitog i postepenog učitavanja rečnika.

C. *Kombinovano učitavanje rečnika*

Kombinovano učitavanje rečnika je kompromisno rešenje koje koristi najbolje odlike prethodna dva pristupa. Kako alat vrši proveru teksta preko korisničkog interfejsa koji se pokreće u internet pregledaču, vreme potrebno za njegovo pokretanje je iskorišćeno za učitavanje rečnika. Tako se vreme koje korisnik mora da sačeka koristi da se u pozadini učita što je veći deo rečnika moguć. U većini slučajeva se kompletan rečnik tokom pokretanja alata učita međutim, u retkim situacijama, konekcija sa bazom će biti prekinuta, te se učitavanje rečnika na to slovo odlaže do trenutka kada je rečnik na to slovo potreban.

Odabrani pristup doprinosi tome da korisnik ima u većini slučajeva potpuno spreman alat za proveru pravopisa, a da u preostalim slučajevima ne čeka na pokretanje alata, već da ima privid da je alat spreman za rad, a da se neočekivane greške otklone tokom rada alata vodeći računa da korisnik to ne primeti.

Kako se učitavanje obavlja u paralelnoj *for* petlji, treba očekivati da se za isto slovo pokrenu dodatne niti koje bi nadomestile prekid učitavanja delova rečnika, pa je potrebno obezbediti otklanjanje utrkivanja niti nad deljenom strukturom rečnika. Da bi se utrkivanje niti sprečilo, koristi se niz od 30 mutex semafora za svako slovo rečnika, koji obezbeđuju da se u slučaju kada je alat pokrenut, a rečnik nije kompletno učitan, dodatne niti koje su pokrenute ne prekinu niti ponište rad niti koja već vrši učitavanje rečnika na to slovo. U slučaju kada je pokrenuta dodatna nit, ona će proveriti da li je semafor slobodan. Ukoliko jeste, ni jedna nit ne vrši učitavanje rečnika, na osnovu čega se zaključuje da rečnik na to slovo ili nije učitan ili je učitavanje

 ISBN 978-86-7466-969-3

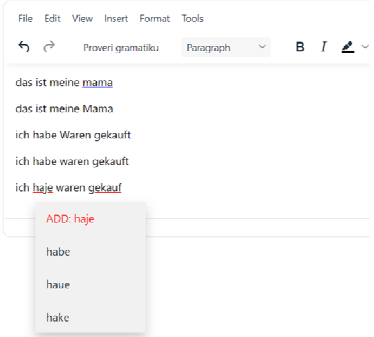
prekinuto, te se učitavanje rečnika na odabrano slovo ponovo pokreće. Ukoliko je semafor zauzet, nova nit nije potrebna iz razloga što već postoji nit koja vrši učitavanje tog slova alfabeta ili je rečnik za to slovo kompletno učitan, te ona prestaje sa radom.

V. KORISNIČKI INTERFEJS ALATA ZA PROVERU PRAVOPISA

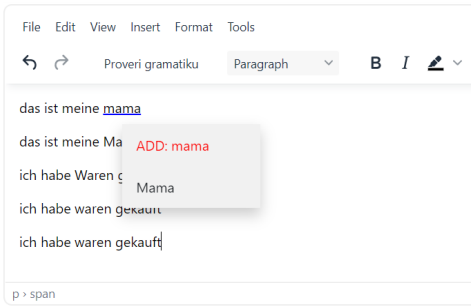
Softverski alat je zamišljen kao internet aplikacija minimalističkog dizajna i intuitivnog korisničkog interfejsa. Za polje za unos teksta iskorišćen je „*Open Source TinyMCE rich HTML text editor*“, koji je prilagođen dodavanjem novih dugmadi za poziv funkcije za proveru sintaksne ispravnosti unetog teksta.

Greške u alatu prikazuju se različitim bojama. Tako su reči koje su neispravno napisane podvučene crvenom bojom (Sl. 6), a plavom bojom su podvučene reči koje su napisane sa pogrešnim početnim slovima (Sl. 7). Sve imenice u nemačkom jeziku, bez obzira na poziciju u rečenici, treba da se pišu velikim početnim slovom. Na primer, rečenica *Das ist meine mama* – ovo je moja mama, iako gramatički tačna, pravopisno nije ispravna, jer reč mama mora biti napisana velikim slovom. Važno je naglasiti da alat nema mogućnost da semantički odredi ispravnost rečenice, te će u slučajevima kada isto napisana reč koja može označavati imenicu ili neku drugu vrstu reči koja se po pravopisu piše malim početnim slovom, smatrati ispravnom bila ona napisana velikim ili malim početnim slovom. Razlog tome je mogućnost alata da u kolekciji reči detektuje unetu reč u onom obliku u kojem je uneta, te je neće smatrati pogrešnom.

Reči nemačkog jezika, kao i u srpskom jeziku mogu imati različita značenja, prema tome mogu pripadati različitim vrstama reči. Tako je korisniku ostavljena mogućnost, da pored podvučenih reči, u rečnik može da doda i nepodvučene reči. Dodavanje reči u rečnik se vrši iz razloga proširenja kolekcije reči radi boljeg rada alata, uz pomoć korisnika koji prilikom odabira opcije dodavanja nove reči mora precizirati njenu vrstu. Proces dodavanja reči u rečnik počinje pritiskom desnog dugmeta miša na samu podvučenu reč, odabirom opcije Add: <reč>, kada se otvara nova lista u kojoj se odabira vrsta reči. Nakon odabira vrste reči, reč će biti upisana u rečnik i odmah se koristi kao referenca za proveru ispravnosti teksta. Vrsta reči je morala da se doda prilikom unošenja nove reči, da bi alat mogao nadalje da vrši sintaksnu analizu rečenice, odnosno proveru ispravnosti redosleda reči.



Sl. 6. Predlaganje ispravljanja greške prouzrokovane neispravnim rečima



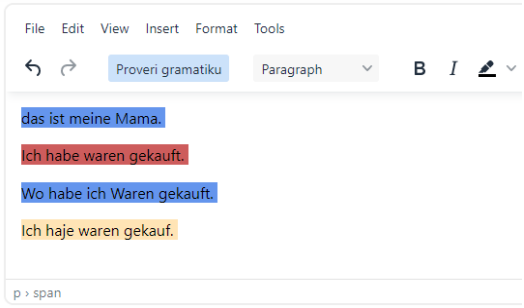
Sl. 7. Predlaganje ispravljanja greške pogrešnog početnog slova

Pod gramatičkom proverom ispravnosti rečenice podrazumeva se sintaksna provera rečenica. Rečenice u nemačkom jeziku mogu biti veoma dugačke, što otežava analizu teksta. Sa druge strane, dugačka rečenica se može posmatrati kao nekoliko kraćih rečenica, kojima je početak nakon nekog od veznika ili znakova interpunkcije. Provera sintaksne ispravnosti vrši se pritiskom na dugme „Proveri gramatiku“. Boje koje su korišćenje u alatu kod ovako izazvanih grešaka su žuta, crvena i plava.

Žutom bojom su označene one rečenice za koje alat ne može da utvrdi ni da su ispravne, ni da su neispravne, iz razloga što u rečenici ili delu rečenice ukoliko je rečenica složena, postoji reč koju alat ne poznaje. To su rečenice koje imaju bar jednu reč koja je podvučena crvenom bojom. Nakon što korisnik doda podvučenu reč u rečnik ili je promeni da bude ispravna, rečenica će biti ponovo proverena.

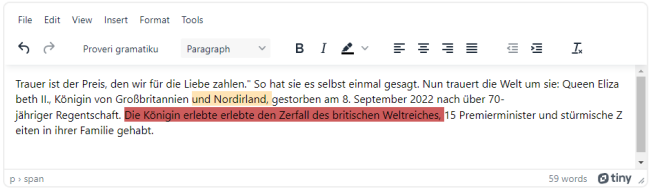
Crvenom bojom su označene rečenice koje su sigurno neispravne. Pod neispravnom rečenicom smatra se ona u kojoj se redosled reči ne uklapa ni u jednu ispravnu formu rečenice.

Ispravna interpunkcija rečenice se smatra preduslovom za ispravnu proveru rečenične sintakse. Zato se plava boja koristi da se naznače rečenice koje su počele malim slovom ili upitne rečenice koje se ne završavaju upitnikom. Izgled označenih rečenica, nakon provere gramatike, dat je na sl. 8.



Sl. 8. Izgled označenih rečenica

Primer analize ispravnosti složenica na odlomku iz novinskog članka iz *Deutsche welle*, prikazan je na slici 9. Rečenica sadrži jednu reč koja je nepoznata, a jedna rečenica je namerno promenjena tako da bude neispravna.



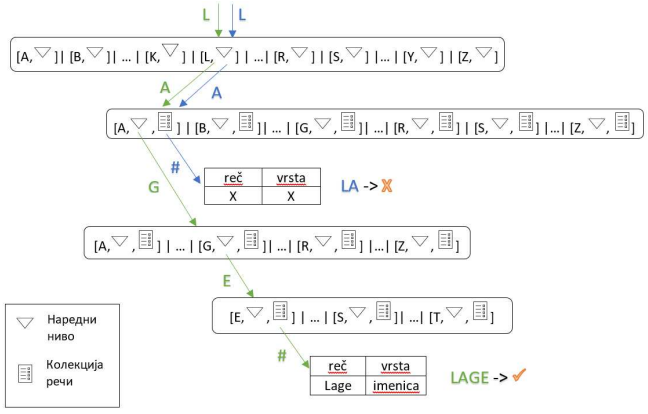
Sl. 9. Izgled provere složenih rečenica

 ISBN 978-86-7466-969-3

U alatu je posebno korisna opcija za osobe koje žele da unaprede znanje nemačkog jezika, jer je moguć pregled reči rečenice i njene određene vrste u rečenici. Nakon što je izvršena sintaksna provera rečenice, korisniku se pruža mogućnost da pritiskom levog tastera miša na neku od rečenica zahteva prikaz vrsta reči u rečenici i time se uveri u ispravnost rečenice i potencijalno je prema tome po želji promeni. Na primer, pregledom vrsta reči rečenice *Ich habe waren gekauft* korisnik će videti da je upotrebio tri glagola zaredom i time zaključiti da je takav oblik rečenice neispravan, te da mu je potrebna imenica *Waren* umesto glagola *waren*. Primer je prikazan na slici 10. Nakon što korisnik bude izvršio korekciju teksta, ponovo će pokrenuti proces provere gramatike, pritiskom na dugme „Proveri gramatiku“, kada će rečenica biti ispravna (*Ich habe Waren gekauft*), a redosled vrsta reči u rečenici će odgovarati poznatoj ispravnoj sintaksi: lična zamenica - glagol - imenica - glagol. Ukoliko se neka reč u rečenici smatra nepoznatom, a pokrenut je proces sintaksne provere teksta, rečenica će biti obojena žutom bojom, a pregled vrsta reči će naznačiti upitnikom vrstu nepoznate reči.

Sl. 10. Pregled vrsta reči u rečenici

Reči nemačkog jezika mogu imati više značenja, a samim tim mogu potpadati pod više vrsta. Alat, iako u tim slučajevima vrši odabir vrste reči prema susednim rečima i ostalim rečima rečenice, može pogrešiti iz razloga što se ne vrši semantička analiza rečenice, već se reči posmatraju prema njihovim vrstama. Kako se može desiti da alat napravi grešku, a samim tim potencijalno ispravnu rečenicu smatra neispravnom iz razloga pogrešno odabrane vrste, korisniku je pružena mogućnost da promeni vrstu reči u rečenici.



Sl. 11. Proces pretrage stabla

Sumirano, reč se smatra ispravnom ako se prolaskom kroz stablo po njenim slovima može doći do terminalnog znaka. Kada se terminalni znak dostigne, reč mora postojati u kolekciji reči njenog krajnjeg slova. Primer rada pretrage strukture u uspešnom i neuspešnom slučaju prikazan je na slici 11. Nakon nalaženja reči, u zavisnosti od njene vrste se određuje da li je ona napisana ispravnim početnim slovom.

VI. Zaključak

Cilj ovog rada bio je razvoj alata za proveru ispravnosti teksta zasnovan na poznatim strukturama podataka. Takođe, cilj je bio razviti veb rešenje, koje radi brzo, pouzdano i koje pruža visoku ugodnost korišćenja. U istraživanju koje je prethodilo razvoju ovog alata, urađene su analize postojećih dostupnih alata. Primećeno je da su alati za sve strane jezike, osim engleskog jezika, znatno manje zastupljeni i razvijani, pa je proces učenja za sve koji se služe takvim jezicima otežan. Alati koji su imali podržan nemački jezik, uglavnom su bili zasnovani na prevodu i semantičkoj analizi teksta, zapostavljajući važnost ispravnosti gramatike. Jedan od većih izazova na razvoju alata bilo je nalaženje pogodnog i obimnog rečnika, sa raznovrsnošću reči, tačnom i dostupnom gramatikom, pa je odabran zvaničan rečnik nemačkog jezika, koji se dalje obrađivao.

U okviru alata je razvijen veliki broj funkcionalnosti koje doprinose zadovoljenju korisničkih potreba kao što su: provera unosa teksta, provera velikog i malog slova, generisanje predloga za neispravne reči i zamena neispravne reči ispravnom iz prozora predloženih reči, provera sintakse i interpunkcije rečenice, uvid u sve vrste reči teksta i ručna promena vrste reči koja će se u rečenici koristiti.

Realizovani alat je otvoren za proširenja i mogućnost implementacije semantičke analize sa ciljem unapređenja analize gramatike. Trenutno implementirani alat nije u stanju da razume smisao rečenice, a slučajeve kada jedna reč može imati više vrsta rešava odabirom neke od vrsta na osnovu susednih reči. Implementacija pomenute semantičke analize je izuzetno zahtevan, obiman i dugotrajan proces, ali koji bi na kraju doveo do toga da alat raspolaže praktično svim funkcionalnostima za potrebe unapređenje znanja ovog jezika.

Zahvalnica

Ovo istraživanje je finansirano od strane Ministarstva nauke, tehnološkog razvoja i inovacija Republike Srbije (ugovor broj: 451-03-47/2023-01/200103) i evropskog projekta Horizon 2020 - *European federation of Data Driven Innovation Hubs*.

Literatura

[1] T. Fitria, “Grammarly as AI-powered English Writing Assistant: Students’ Alternative for English Writing,” Metathesis: Journal of English language, literature and teaching, vol. 5, no. 1, pp. 65-78, 2021.

[2] A. Hering, M.Matussek, Michaela Perlmann-Balme, “*Deutsch Übungsgrammatik für die Mittelstufe aktuell*,” 1st ed. Hueber Verlag, 2022.

[3] Duden, „Deutsches Universalwörterbuch,“ 8th ed, Bibliographisches Institut, 2015.

[4] PONS internet dictionary and translator, dostupno na: https://en.pons.com/translate/german-english/unterbringen (1.3.2023.)

[5] D.Drašković, N.Kojić, M.Mićović, U.Radenković, “Implementacija sistema za prikupljanje podataka, generisanje klastera i preporuka pomoću mašinskog učenja,” Zbornik radova 24. konferencije YU INFO 2018, pp. 167-172, Kopaonik, Srbija, mart 2018.

[6] S. Koltovskaia, “Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study,” Assessing Writing, vol. 44, 2020.

[7] Insta Text, dostupno na: https://instatext.io/ (22.4.2023.)

 ISBN 978-86-7466-969-3