LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

OCR pielāgošana vairāku alfabētu tekstiem

BAKALAURA DARBS

Autors: **Roberts Brants**
Studenta apliecības Nr.: rb19043
Darba vadītājs: Dr. dat. Pēteris Paikens

RĪGA 2023

ANOTĀCIJA

Bakalaura darba mērķis ir izpētīt metodes, kas palielina OCR tehnoloģijas precizitāti dažādu alfabētu tekstu attēliem. Kā arī, izstrādāt tādu OCR adapteri, kas sekmīgi implementē izpētītās metodes, un spēj apstrādāt dažādu alfabētu rakstzīmju saturošu tekstu. Šāds adapteris nepieciešams, lai sekmētu sarežģīta satura tekstu digitalizēšanas procesu, tādejādi sekmējot algoritmisku procesu pielietojamību digitalizētiem tekstiem. Darba pamatā tiek izmatots Konstantīna Karuļa "Latviešu etimoloģijas vārdnīca" (Avots, 2001) grāmatas LNB digitalizējums.

Atslēgvārdi: OCR, attēli, rakstzīmju atpazīšana.

ABSTRACT

“Adapting optical character recognition (OCR) to multi-alphabet texts”.

The aim of the bachelor's thesis is to study methods that increase the accuracy of OCR technology for text images of different alphabets. Also, develop an OCR adapter that successfully implements the studied methods and can process text containing characters of various alphabets. Such an adapter is necessary to facilitate the digitization process of texts with complex content, thereby facilitating the applicability of algorithmic processes to digitized texts. The work is based on the LNB digitization of the book "Latviešu etimoloģijas vārdnīca" (Avots, 2001) by Konstantins Karulis.

Keywords: OCR, images, character recognition.

SATURS

APZĪMĒJUMU SARAKSTS

OCR (*Optical Character Recognition*) - optiskā rakstzīmju atpazīšana attēlā.

LNB – Latvijas Nacionālā bibliotēka.

WER (*Word Error Rate*) – vārdu kļūdu līmenis

CER (*Character Error Rate*) - rakstzīmju līmenis

LNB – Latvijas Nacionālā bibliotēka.

IEVADS

OCR tehnoloģija ir programmatūras risinājums, kurš spēj digitalizēt tekstu no tekstu saturošiem attēliem. Šī tehnoloģija tiek plaši pielietota ikdienā - automobiļu numuru atpazīšanai, grāmatu digitalizēšanai, dokumentu pārvaldībai, un daudzkur citur. Tādēļ ir svarīgi, lai OCR digitalizētais produkts būtu precīzs un pareizs ne tikai triviālos gadījumos, bet arī gadījumos, kad teksts ir sarežģīts un satur vairāk kā viena alfabēta burtus, piemēram Konstantīna Karuļa (1915.-1997.) Latviešu etimoloģijas vārdnīca.

Līdz ar to darba mērķis ir izzināt un eksperimentālā ceļā analizēt metodes, kas sekmē OCR rezultātu pareizību dažādu alfabētu tekstiem. Darba pamatā tiek izmantota Roberta Branta kursa darbs (2023) "Optiskas rakstzīmju pazīšanas (OCR) tehnoloģiju pielietojums latviešu etimoloģijas vārdnīcām."

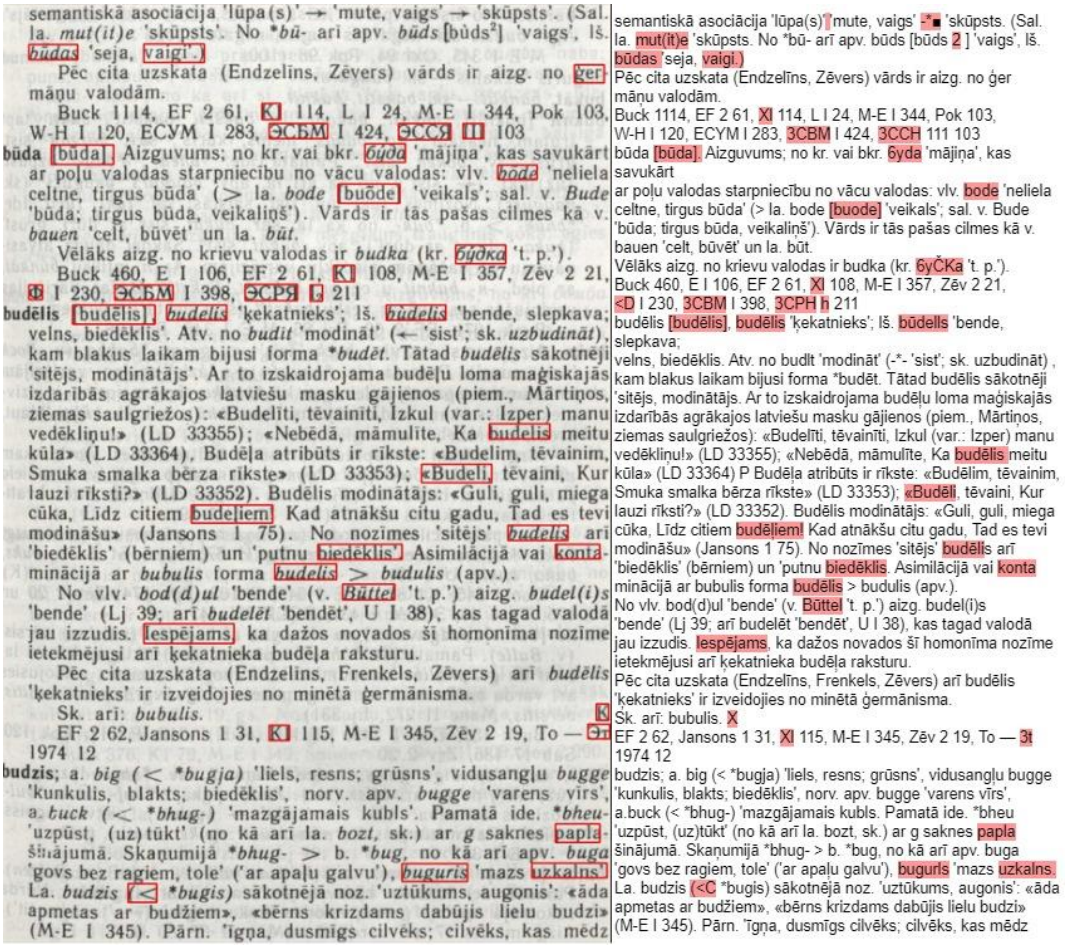
Darba gaitā tiek izvirzīti sekojoši uzdevumi:

- izzināt OCR tehnoloģiju un pieejamos risinājumus;
- izzināt teksta līdzības metrikas, kuras var pielietot OCR rezultātu analīzei;
- izpētīt un analizēt kā segmentēt teksta apgabalus;
- eksperimentālā ceļā izpētīt kā sekmēt OCR precizitāti segmentētajiem tekstu apgabaliem;
- izstrādāt tādu OCR rīku adapteri, kas spēj atpazīt tekstu attēlā ar augstu precizitāti.

1. SITUĀCIJAS RAKSTUROJUMS

Tirgū ir pieejami dažādi OCR risinājumi, taču tie lielākoties ir mērķēti tekstu attēliem kuros tiek attēlots viena veida alfabēta rakstzīmes no vienas valodas. Daži no šiem rīkiem piedāvā arī valodas atpazīšanas funkcionalitāti vai pat vairāku valodu konfigurācijas, taču šīs programmatūras tiek stingri ierobežotas tiem pieejamajās valodās, tādejādi nespējot dinamiski adaptēties komplicēta satura tekstam, kas satur vairāku valodu, vairāku alfabētu rakstzīmes.

Latvijas Nacionālā Bibliotēka (LNB) aktīvi veic savu grāmatu krājuma digitalizāciju, izmantojot *Abbyy Finereader* un *Tesseract* rīkus. Šī digitalizēšanas procesa produkti ir publiski pieejami LNB digitālajā bibliotēkā [3]. Liela daļa šo digitālo publikāciju, pēc autora domām, ir pareiza, lasāma, uztverama (no digitalizēšanas rezultātu viedokļa). Taču ir arī daļa komplicētāka satura publikāciju, kuru digitālais produkts gala patērētājam apgrūtina teksta uztveri un analīzi, piemēram (sk. att. 1.1) Konstantīna Karuļa (1915.-1997.) *Latviešu etimoloģijas vārdnīca* (Avots, 2001), kas ir digitalizēta LNB 2018. gadā [2], atjaunota 2023. gadā.


semantiskā asociācija 'lūpa(s)' → 'mute, vaigs' → 'skūpsts'. (Sal. la. *mut(it)e* 'skūpsts'. No **bū-* arī apv. *būds* [būds²] 'vaigs', lš. *būdas* 'seja, vaigi'.)
Pēc cita uzskata (Endzelīns, Zēvers) vārds ir aizg. no ģermāņu valodām.
Buck 1114, EF 2 61, KI 114, L I 24, M-E I 344, Pok 103, W-H I 120, ECУМ I 283, ЭСБМ I 424, ЭССЯ III 103
būda [būda]. Aizguvums; no kr. vai bkr. *бýда* 'mājiņa', kas savukārt ar poļu valodas starpniecību no vācu valodas: vlv. *bōde* 'neliela celtne, tirgus būda' (> la. *bode* [buōde] 'veikals'; sal. v. *Bude* 'būda; tirgus būda, veikaliņš'). Vārds ir tās pašas cilmes kā v. *bauen* 'celt, būvēt' un la. *būt*.
Vēlāks aizg. no krievu valodas ir *budka* (kr. *бýдка* 't. p.').
Buck 460, E I 106, EF 2 61, KI 108, M-E I 357, Zēv 2 21, ФВ I 230, ЭСБМ I 398, ЭСРЯ I 211
budēlis [budēlis]. *budelis* 'ķekatnieks'; lš. *būdelis* 'bende, slepkava; velns, biedēklis'. Atv. no *budit* 'modināt' (← 'sist'; sk. *uzbudināt*), kam blakus laikam bijusi forma **budēt*. Tātad *budēlis* sākotnēji 'sitējs, modinātājs'. Ar to izskaidrojama budēļu loma maģiskajās izdarībās agrākajos latviešu masku gājienos (piem., Mārtiņos, ziemas saulgriežos): «Budelīti, tēvainīti, Izkul (var.: Izper) manu vedēkliņu!» (LD 33355); «Nebēdā, māmulīte, Ka budelis meitu kūla» (LD 33364). Budēļa atribūts ir rīkste: «Budelim, tēvainim, Smuka smalka bērza rīkste» (LD 33353); «Budeli, tēvaini, Kur lauzi rīksti?» (LD 33352). Budēlis modinātājs: «Guli, guli, miega cūka, Līdz citiem budeļiem. Kad atnākšu citu gadu, Tad es tevi modināšu» (Jansons I 75). No nozīmes 'sitējs' *budelis* arī 'biedēklis' (bērniem) un 'putnu biedēklis'. Asimilācijā vai kontaminācijā ar *bubulis* forma *budelis* > *budulis* (apv.).
No vlv. *bod(d)ul* 'bende' (v. *Büttel* 't. p.') aizg. *budel(i)s* 'bende' (Lj 39; arī *budelēt* 'bendēt', U I 38), kas tagad valodā jau izzudis. Iespējams, ka dažos novados šī homonīma nozīme ietekmējusi arī ķekatnieka budēļa raksturu.
Pēc cita uzskata (Endzelīns, Frenkels, Zēvers) arī *budēlis* 'ķekatnieks' ir izveidojies no minētā ģermānisma.
Sk. arī: *bubulis*.
EF 2 62, Jansons I 31, KI 115, M-E I 345, Zēv 2 19, To — Etn 1974 12
budzis; a. *big* (< **bugja*) 'liels, resns; grūsns', vidusangļu *bugge* 'kunkulis, blakts; biedēklis', norv. apv. *bugge* 'varens vīrs', a. *buck* (< **bhug-*) 'mazgājamais kubls'. Pamatā ide. **bheu-* 'uzpūst, (uz)tūkt' (no kā arī la. *bozt*, sk.) ar *g* saknes paplašinājumā. Skaņumijā **bhug-* > b. **bug*, no kā arī apv. *buga* 'govs bez ragiem, tole' ('ar apaļu galvu'), *buguris* 'mazs uzkalns'. La. *budzis* (< **bugis*) sākotnējā noz. 'uztūkums, augonis': «āda apmetas ar budžiem», «bērns krizdams dabūjis lielu budzi» (M-E I 345). Pārn. 'īgņa, dusmīgs cilvēks; cilvēks, kas mēdz

semantiskā asociācija 'lūpa(s)' mute, vaigs' -* skūpsts. (Sal. la. mut(it)e 'skūpsts. No *bū- arī apv. būds [būds 2] 'vaigs', lš. būdas 'seja, vaigi.)
Pēc cita uzskata (Endzelīns, Zēvers) vārds ir aizg. no ģer māņu valodām.
Buck 1114, EF 2 61, XI 114, L I 24, M-E I 344, Pok 103, W-H I 120, ECYM I 283, 3CBM I 424, 3CCH 111 103
būda [būda]. Aizguvums; no kr. vai bkr. 6yda 'mājiņa', kas savukārt
ar poļu valodas starpniecību no vācu valodas: vlv. bode 'neliela celtne, tirgus būda' (> la. bode [buode] 'veikals'; sal. v. Bude 'būda; tirgus būda, veikaliņš'). Vārds ir tās pašas cilmes kā v. bauen 'celt, būvēt' un la. būt.
Vēlāks aizg. no krievu valodas ir budka (kr. 6yČKa 't. p.').
Buck 460, E I 106, EF 2 61, XI 108, M-E I 357, Zēv 2 21, <D I 230, 3CBM I 398, 3CPH h 211
budēlis [budēlis], budēlis 'ķekatnieks'; lš. būdells 'bende, slepkava;
velns, biedēklis. Atv. no budlt 'modināt' (-*- 'sist'; sk. uzbudināt) , kam blakus laikam bijusi forma *budēt. Tātad budēlis sākotnēji 'sitējs, modinātājs. Ar to izskaidrojama budēļu loma maģiskajās izdarībās agrākajos latviešu masku gājienos (piem., Mārtiņos, ziemas saulgriežos): «Budelīti, tēvainīti, Izkul (var.: Izper) manu vedēkliņu!» (LD 33355); «Nebēdā, māmulīte, Ka budēlis meitu kūla» (LD 33364) P Budēļa atribūts ir rīkste: «Budēlim, tēvainim, Smuka smalka bērza rīkste» (LD 33353); «Budēli, tēvaini, Kur lauzi rīksti?» (LD 33352). Budēlis modinātājs: «Guli, guli, miega cūka, Līdz citiem budēļiem! Kad atnākšu citu gadu, Tad es tevi modināšu» (Jansons 1 75). No nozīmes 'sitējs' budēlis arī 'biedēklis' (bērniem) un 'putnu biedēklis. Asimilācijā vai konta minācijā ar bubulis forma budēlis > budulis (apv.).
No vlv. bod(d)ul 'bende' (v. Büttel 't. p.') aizg. budel(i)s 'bende' (Lj 39; arī budelēt 'bendēt', U I 38), kas tagad valodā jau izzudis. Iespējams, ka dažos novados šī homonīma nozīme ietekmējusi arī ķekatnieka budēļa raksturu.
Pēc cita uzskata (Endzelīns, Frenkels, Zēvers) arī budēlis 'ķekatnieks' ir izveidojies no minētā ģermānisma.
Sk. arī: bubulis. X
EF 2 62, Jansons 1 31, XI 115, M-E I 345, Zēv 2 19, To — 3t 1974 12
budzis; a. big (< *bugja) 'liels, resns; grūsns', vidusangļu bugge 'kunkulis, blakts; biedēklis', norv. apv. bugge 'varens vīrs', a.buck (< *bhug-) 'mazgājamais kubls. Pamatā ide. *bheu 'uzpūst, (uz)tūkt' (no kā arī la. bozt, sk.) ar g saknes papla šinājumā. Skaņumijā *bhug- > b. *bug, no kā arī apv. buga 'govs bez ragiem, tole' ('ar apaļu galvu'), buguris 'mazs uzkalns. La. budzis (<C *bugis) sākotnējā noz. 'uztūkums, augonis': «āda apmetas ar budžiem», «bērns krizdams dabūjis lielu budzi» (M-E I 345). Pārn. 'īgņa, dusmīgs cilvēks; cilvēks, kas mēdz

Att. 1.1. Karulis, Konstantīns, Latviešu etimoloģijas vārdnīca. Rīga: Avots, 2001. 151. lpp. LNB digitalizējums (kreisajā pusē oriģinālais attēls, labajā pusē digitalizācijas produkts), sarkanā krāsā: kļūdas.

No veiktā pētījuma [1] par K. Karuļa *Latviešu etimoloģijas vārdnīca* digitalizēšanu var secināt, ka *Abbyy Finereader* rīks spēj precīzāk atpazīt komplicēta satura teksta attēlus nekā *Tesseract* rīks, taču *Tesseract* spēj tikpat labi atpazīt sarežģīta satura teksta attēlus kā *Abbyy Finereader* pie *Tesseract* vārdu pārliecinātības sliekšņa 35% (latviešu valodas konfigurācijā). Darbā tika arī identificētas un analizētas rīku pieļautās kļūdas, kā piemēram: problēmas efektīvi pārslēgties no vienas valodas alfabēta uz citas valodas alfabētu, pareizi lietot, atpazīt pēdiņas un nespēja pārliecinoši atpazīt līdzīgas formas rakstzīmes.

Šīs problēmas ir iespējams risināt izstrādājot OCR rīka pēcapstrādes algoritmu, kas izmanto etimoloģijas vārdnīcas struktūru, kas definēta vārdnīcā. Šāda pieeja ļauj apstrādāt vārdus vai teikumus ar īpašu nozīmi, piemēram, vārdu saīsinājumi, kas apzīmē nākamā vārda svešvalodu, kvadrātiekavas, kas apzīmē iepriekšējā vārda izrunu, vai rindkopas, kur zināms, ka pēdējā rindkopa vienmēr ir paredzēta atsaucēm.

2. OPTISKĀ RAKSTZĪMJU ATPAZĪŠANA (OCR)

OCR ir tehnoloģija, kas pārvērš tekstu no skenētiem dokumentiem, attēliem vai citiem nedigitāliem avotiem par digitālu tekstu, ko var apstrādāt un analizēt datorā. Šī tehnoloģija ir fundamentāla datu ieguves, dokumentu digitalizācijas un automātiskās informācijas apstrādes procesos.

OCR tehnoloģijas attīstība aizsākās 20. gadsimta vidū, taču tās pilnība un precizitāte ir būtiski uzlabojusies pēdējās desmitgadēs, galvenokārt datoru redzes un mašīnmācīšanās tehnoloģiju attīstības dēļ. Modernās OCR sistēmas ir spējīgas atpazīt dažādu veidu fontus un rakstības, tostarp pat roku rakstītu tekstu, ar aizvien augstāku precizitāti.

OCR tehnoloģija spēj atpazīt tekstā esošās rakstzīmes un pārvērst tās digitālā formātā, piemēram, teksta failā vai datu bāzē, kas pēc tam var tikt apstrādāts un analizēts. Šī tehnoloģija ir ļoti noderīga daudzās jomās, piemēram, dokumentu digitalizācijā, datorizētā redzē, automātiskajā datu ievadē, teksta analīzē un citās.

OCR tehnoloģijas izmantošana ir plaši izplatīta daudzās industrijās un nozarēs. Piemēram, bankās un finanšu iestādēs, lai automātiski apstrādātu čekus un rēķinus, bibliotēkās un arhīvos, lai digitalizētu vecus dokumentus un grāmatas, pakalpojumu sniedzējiem, lai automātiski ievadītu datus no formām un dokumentiem, un tā tālāk.

2.1. OCR tehnoloģijas darbība

OCR darbības princips ietver vairākus soļus. Vispirms tiek veikta attēla priekšapstrāde, lai uzlabotu tā kvalitāti un atvieglotu rakstzīmju atpazīšanu. Tas ietver attēla mērogošanu, trokšņu samazināšanu, kontrasta uzlabošanu un citus uzlabojumus. Tad seko segmentācija, kurā attēls tiek sadalīts atsevišķās rakstzīmēs vai vārdos. Tad seko rakstzīmju atpazīšana, izmantojot mašīnmācīšanās modeļus. Beigās teksts tiek pārbaudīts un koriģēts, izmantojot vārdnīcas un gramatikas pārbaudi.

2.1.1. Attēlu priekšapstrāde

OCR tehnoloģija sākas ar procesu, kas tiek saukts par attēlu priekšapstrādi. Šis process ir būtisks, lai uzlabotu attēla kvalitāti un vienkāršotu rakstzīmju atpazīšanu. OCR attēlu priekšapstrāde var ietvert vairākus soļus: krāsu normalizācija un binarizācija, trokšņu noņemšana, skalēšana un rotācija, .

Krāsu normalizācija tiek izmantota, lai samazinātu attēlā esošo krāsu skaitu, kas atvieglo turpmāko apstrādi. Tas bieži vien tiek izdarīts, pārveidojot attēlu melnbaltā formātā. Binarizācija ir process, kurā attēls tiek pārveidots, lai tas saturētu tikai divus toņus - melno un

balto. Tas tiek darīts, izmantojot sliekšņošanas tehniku, kurā tiek noteikts intensitātes slieksnis, un visi pikseļi, kas ir virs šī sliekšņa, tiek uzskatīti par balto, bet tie, kas ir zemāki, tiek uzskatīti par melno. Tas palīdz izcelt teksta rakstzīmes un atdalīt tās no fona.

Trokšņu noņemšana ir svarīgs priekšapstrādes solis, kas palīdz samazināt attēla kvalitātes zudumus, kas var rasties nevēlamu izmaiņu dēļ attēla kvalitātē, piemēram, pikseļu izplūšanas, skenēšanas kļūdas, u.c. Trokšņu noņemšanai tiek izmantotas dažādas attēlu apstrādes metodes, tostarp filtrēšana (piemēram, vidējā vērtība, medians), slīpēšana un morfoloģiskās transformācijas.

Skalēšana nodrošina, ka visi attēli tiek apstrādāti vienotā izmērā, kas palīdz normalizēt OCR procesu. Turklāt, rotācija tiek veikta, lai nodrošinātu, ka teksts ir pareizā orientācijā. Teksta orientācija ir būtiska OCR precizitātei, jo OCR algoritmi parasti ir optimizēti strādāt ar horizontālu tekstu. Ja teksts ir pagriezts, tas var traucēt rakstzīmju atpazīšanu.

2.1.2. Segmentācija

Segmentācija ir būtisks solis optiskās rakstzīmju atpazīšanas procesā. Šajā solī attēls tiek sadalīts atsevišķos segmentos, kas atbilst atsevišķiem rakstiem vai vārdiem. Segmentācijas mērķis ir nodrošināt, ka katru rakstzīmi var atpazīt atsevišķi. Attēlu segmentācija var ietvert vairākus soļus: bloku segmentācija, rindiņu segmentācija, vārdu segmentācija, rakstzīmju segmentācija.

Bloku segmentācija ir pirmā līmeņa segmentācijas solis, kurā tiek noteikti teksta bloki attēlā. Katrs bloks satur fragmentu no kopējā teksta. Attēliem ar strukturētu tekstu, piemēram, grāmatas lappusēm, šo soli var viegli realizēt, izmantojot vienkāršus algoritmus, kas balstīti uz tekstā esošo telpisko izkārtojumu. Taču sarežģītākiem attēliem, piemēram, ceļa zīmēm vai brošūrām, var būt nepieciešama sarežģītāka pieeja, ieskaitot mašīnmācīšanās metodes.

Kad ir noteikti teksta bloki, nākamais solis ir rindiņu segmentācija. Šajā solī katrai rindiņai attēlā tiek piešķirts atsevišķs segmentēšanas identifikators. Tas palīdz atpazīšanas procesā, jo rakstzīmes vienā rindiņā parasti ir līdzīgas orientācijas un izmēra. Šī procesa izaicinājums ir atrast pareizo līniju izvietojumu, kas var būt īpaši grūti attēliem ar slīpām līnijām vai rakstzīmēm.

Rindiņu segmentācija tiek papildināta ar vārdu segmentāciju. Šajā solī tiek veikta līdzīga procedūra kā rindiņu segmentācijā, taču šoreiz tiek noteikti atsevišķi vārdi katrā rindiņā. Šis solis palīdz atpazīšanas procesā, jo bieži vien rakstzīmes vienā vārdā ir līdzīgas. Tomēr šis solis var būt izaicinājums, jo atstarpes starp vārdiem var būt dažādas.

Pēdējais segmentācijas solis ir rakstzīmju segmentācija. Šajā solī tiek noteiktas atsevišķas rakstzīmes katrā vārdā. Rakstzīmju segmentācija ir būtiska, lai nodrošinātu precīzu

rakstzīmju atpazīšanu. Šis solis var būt izaicinājums, jo dažās situācijās rakstzīmes var būt cieši pievienotas vai pat pārklājas. Tādēļ šis solis var prasīt īpaši precīzas un rafinētas metodes, ieskaitot mašīnmācīšanās algoritmus.

2.1.3. Rakstzīmju atpazīšana

Pēc attēla priekšapstrādes un segmentācijas nākamais OCR process ir rakstzīmju atpazīšana. Šajā posmā sistēma mēģina identificēt un klasificēt katru atsevišķo rakstzīmi. Rakstzīmju atpazīšanas process ietver šādus posmus: rakstzīmju īpašību izgūšana, rakstzīmju klasifikācija, atzīšana un interpretācija.

Rakstzīmju īpašību izgūšanas procesā no katras rakstzīmes tiek izgūtas noteiktas īpašības, kas palīdz to identificēt. Šīs īpašības var ietvert dažādus parametrus, piemēram, rakstzīmes formu, izmēru, kontūru, tekstūru un citus vizuālos elementus. Īpašību izgūšana ir būtiska, jo tā nodrošina, ka katru rakstzīmi var precīzi un efektīvi salīdzināt ar zināmu rakstzīmju kopumu.

Rakstzīmju klasifikācijas posmā tiek izmantoti mašīnmācīšanās algoritmi, lai "iemācītos" atpazīt rakstzīmes. Šie algoritmi bieži vien tiek apmācīti, izmantojot lielas rakstzīmju datu bāzes, un tiek optimizēti, lai identificētu rakstzīmes ar pēc iespējas lielāku precizitāti. Rakstzīmju klasifikācijas algoritmi var būt vienkārši (piemēram, *k-nearest neighbours*), bet bieži vien tie ir diezgan sarežģīti un ietver tādus elementus kā konvolūcijas neironu tīkli (CNN), kas ir īpaši labi piemēroti vizuālu īpašību apstrādei.

Pēc tam, kad rakstzīmes ir klasificētas, tās tiek interpretētas kā teksts. Šajā solī var tikt izmantota konteksta analīze, lai palīdzētu atpazīt rakstzīmes, kas var būt grūti atpazīstamas vai neviennozīmīgas. Piemēram, ja OCR sistēma nespēj precīzi identificēt konkrētu rakstzīmi, tā var izmantot konteksta informāciju (piemēram, apkārtējos vārdus), lai noteiktu, kura rakstzīme ir visvēršamākā.

2.1.4. Pēcapstrāde

OCR pēcapstrāde ir svarīgs process, kas palīdz uzlabot galīgo OCR rezultātu kvalitāti. Tas ietver vairākus soļus, kuros tiek labots un interpretēts OCR sistēmas iegūtais teksts. Šie ir daži no galvenajiem OCR pēcapstrādes posmiem: pareizrakstības pārbaude, konteksta analīze, datu strukturēšana un formātu pārveide un manuāla korekcija.

Pareizrakstības pārbaudes posms ir ļoti svarīgs, lai labotu OCR rezultātus. Pareizrakstības pārbaude pārbauda iegūtā teksta atbilstību pareizrakstības vārdnīcām, lai identificētu un labotu iespējamas rakstzīmju atpazīšanas kļūdas. Vārdi, kuri nav atrodami vārdnīcā, tiek uzskatīti par iespējamiem kļūdaini atpazītiem. OCR sistēma var izmantot

vairākas stratēģijas, lai labotu šīs kļūdas, piemēram, izmantojot kļūdaino rakstzīmju aizstāšanu ar līdzīgām rakstzīmēm, kas rezultātā veido pareizi rakstītu vārdu.

Konteksta analīze ir paplašināta pareizrakstības pārbaude, kas ne tikai pārbauda vārdu pareizrakstību, bet arī to, kā tie iederas teksta kontekstā. Konteksta analīze var palīdzēt atrisināt daudzas OCR kļūdas, piemēram, palīdzēt saprast, vai teksta fragmentā ir jābūt vārdam "māja" vai "māsa", balstoties uz apkārtējo vārdu secību un nozīmi.

Datu strukturēšanas un formātu pārveides pēcapstrādes posms nodrošina, ka OCR iegūtais teksts tiek pareizi strukturēts un formatēts. OCR var konvertēt tekstu no grafiskā formāta uz mašīnlasāmu formātu, piemēram, PDF vai Word dokumentu. Turklāt, ja sākotnējais dokumentā bija tabulas, diagrammas vai citi strukturēti dati, OCR sistēma var mēģināt atjaunot šo struktūru, lai teksts būtu viegli lasāms un analizējams.

Lai gan OCR tehnoloģijas ir ļoti attīstītas, dažreiz ir nepieciešama manuāla intervence, lai nodrošinātu augstāko iespējamo precizitāti. Šis process ietver cilvēka redaktora vai pārbaudītāja darbu, kurš pārskata OCR rezultātus un labo kļūdas, kuras OCR sistēma varēja palaist garām. Lai gan šis process var būt laikietilpīgs, tas var būt nepieciešams situācijās, kad ir vajadzīga augsta precizitāte, piemēram, juridiskos dokumentos vai citos oficiālos dokumentos.

2.2. Pieejamie OCR rīki

OCR tehnoloģijas ir plaši pieejamas un ietver dažādas formās. Ir gan bezmaksas atvērtā koda rīki, gan komerciāli risinājumi, kas integrējas ar mākoņpakalpojumiem un biznesa aplikācijām. Tas ļauj izvēlēties piemērotu risinājumu atkarībā no individuālajām vajadzībām un budžeta.

Tesseract ir viens no vadošajiem bezmaksas OCR dzinējiem, ko sākotnēji izstrādāja *HP* un tagad uztur *Google*. Tesseract piedāvā augstu atpazīšanas precizitāti un atbalsta vairāk nekā 100 valodas. Tā ir bibliotēka, kas piedāvā OCR funkcionalitāti un to var integrēt dažādās programmēšanas valodās.

Abbyy FineReader ir vadošais komerciālais OCR rīks, kas piedāvā plašas funkcijas, ieskaitot OCR, dokumentu salīdzināšanu, PDF rediģēšanu un citas funkcijas. Tas atbalsta vairāk nekā 190 valodas un spēj apstrādāt dažādus dokumentu formātus.

Bez maksas pieejamie rīki, piemēram, *Tesseract* un *OCR.space,* sniedz pamata OCR funkcionalitāti un atbalsta vairākas valodas. Tie ir īpaši noderīgi mazāk sarežģītos un budžeta ierobežojumus diktējošos projektos.

Komerciālie risinājumi, piemēram, *Abbyy FineReader, Adobe Acrobat Pro DC* un *Amazon Textract*, piedāvā plašākas iespējas, precizitāti un integrāciju ar citām

lietojumprogrammām. Tie ir ideāli lielāku projektu vai uzņēmumu vajadzībām, kas prasa augstu precizitāti un papildu funkcionalitāti, piemēram, dokumentu salīdzināšanu vai strukturētas informācijas izgūšanu.

Darba izstrādes procesā autors nolēma izmantot *Tesseract* OCR rīku, un šāda izvēle bija pamatota vairākiem faktoriem. Pirmkārt, *Tesseract* OCR ir bezmaksas rīks, kas nozīmē, ka tas ir pieejams ikvienam bez papildu izmaksām. Tas bija svarīgs aspekts, jo tas ļāva autoram ietaupīt izmaksas un izmantot resursus citos darba izstrādes posmos.

Turklāt *Tesseract* OCR ir populārākais bezmaksas OCR rīks, kas ieguvis plašu atbalstu kopienā. Šī popularitāte sniedz priekšrocības, jo nozīmīgs lietotāju skaits nozīmē, ka ir pieejami plaši resursi, dokumentācija un pieredze, kas var noderēt darba izstrādes procesā. Autoram bija svarīgi strādāt ar rīku, kas ir labi atbalstīts un izmantojams lietotāju kopienā, nodrošinot pieejamību pie kvalitatīviem resursiem un atbalstu gadījumā, ja rodas jautājumi vai problēmas.

3. TEKSTA LĪDZĪBAS METRIKAS

OCR rezultātu precizitātes validācija ir ne tikai būtiska, bet arī nepieciešama, lai nodrošinātu digitālā teksta augstas kvalitātes nodrošināšanu. Šāda validācija ir īpaši svarīga, jo tā palīdz novērtēt, cik efektīvi un precīzi OCR sistēma ir spējusi pārvērst attēlu par tekstu, kas ir viegli lasāms un interpretējams. Daudzos gadījumos OCR tehnoloģijas izmantošanas efektivitātes noteikšana ir tieši saistīta ar tās spēju veikt precīzu teksta atpazīšanu. Nevienā OCR sistēmā nav pilnības, un kļūdas var rasties dažādu iemeslu dēļ, tostarp attēlu kvalitātes, fonta veida un izmēra, teksta izkārtojuma vai pat OCR sistēmas ierobežojumu dēļ.

Lai noteiktu OCR tehnoloģijas veiktspēju un precizitāti, tiek izmantotas validācijas metrikas. Šīs metrikas ir radītas tā, lai kvantitatīvi novērtētu OCR rezultātus, sniedzot skaidrus, izmērāmus rādītājus par to, cik labi OCR sistēma darbojas. OCR validācijas metrikas parasti tiek sadalītas divās galvenajās kategorijās: kvantitatīvās un kvalitatīvās metrikas. Kvantitatīvās metrikas mēra OCR precizitāti, izmantojot skaitliskus rādītājus, piemēram, rakstzīmju kļūdu līmeni (CER) un vārdu kļūdu līmeni (WER). Kvalitatīvās metrikas, no otras puses, novērtē OCR izvades kvalitāti no lietotāja perspektīvas, mērot faktorus, piemēram, teksta salasāmība un semantisko integritāti. Abas metriku kategorijas ir svarīgas, lai pilnībā novērtētu OCR tehnoloģijas veiktspēju un efektivitāti.

3.1. Levenšteina attālums

Levenšteina attālums, pazīstams arī kā rediģēšanas attālums, ir kvantitatīva metrika, kas tiek izmantota, lai novērtētu OCR rezultātu precizitāti. Šī metrika kvantitatīvi novērtē, cik daudz OCR rezultāts atšķiras no oriģinālā dokumenta. Šīs atšķirības tiek mērotas, skaitot minimālo operāciju skaitu, kas nepieciešams, lai pārveidotu OCR izvadīto tekstu par oriģinālo tekstu. Šīs operācijas var ietvert rakstzīmju ievietošanu, dzēšanu vai aizstāšanu.

Levenšteina attāluma aprēķināšanas process ir šāds:

- izveido divdimensiju matricu, kurā viena dimensija pārstāv OCR izvadīto tekstu, bet otra dimensija pārstāv oriģinālo tekstu. Katru rakstzīmi no abiem tekstiem apzīmē kā vienu matricas punktu;
- aizpilda matricu, aprēķinot minimālo operāciju skaitu, kas nepieciešams, lai pārveidotu vienu teksta fragmentu par otru. Katrā matricas punktā saglabā minimālo operāciju skaitu, kas nepieciešams, lai pārveidotu attiecīgo teksta fragmentu;

- beigās, matricas apakšējā labajā stūrī iegūst minimālo operāciju skaitu, kas nepieciešams, lai pārveidotu pilnīgi visu OCR izvadīto tekstu par oriģinālo tekstu. Šis skaitlis ir Levenšteina attālums starp diviem tekstiem.

Levenšteina attālumu bieži izmanto, lai aprēķinātu rakstzīmju kļūdu līmeni (CER), kas ir viena no visizplatītākajām OCR precizitātes metrikām. Tomēr Levenšteina attālums var būt noderīgs arī, lai novērtētu OCR sistēmas veiktspēju vispārīgi, jo tas sniedz skaitlisku mērījumu OCR rezultātu un oriģinālo dokumentu atšķirībai.

3.2. Rakstzīmju vai vārdu kļūdu līmenis (CER/WER)

Rakstzīmju vai vārdu kļūdu līmenis ir viena no visizplatītākajām kvantitatīvajām OCR validācijas tehnikām. Tā prasa salīdzināt OCR rezultātu ar "patieso" tekstu (manuāli transkribētu un pārbaudītu) un skaitīt atšķirību skaitu rakstzīmēs vai vārdos.

Rakstzīmju kļūdu līmenis (CER - Character Error Rate) novērtē, cik liels procentuālais daudzums rakstzīmju OCR rezultātā ir nepareizs, salīdzinot ar oriģinālo tekstu, kas tiek uzskatīts par "patiesību" vai referenci. Lai aprēķinātu CER, tiek izmantots Levenšteina attālums jeb rediģēšanas attālums, kas nosaka minimālo operāciju skaitu (ieskaitot ievietošanu, dzēšanu un aizstāšanu), kas nepieciešams, lai pārveidotu OCR rezultātu par oriģinālo tekstu. CER tiek aprēķināts, dalot šo operāciju skaitu ar oriģinālā teksta rakstzīmju skaitu, un rezultāts tiek izteikts procentos.

Vārdu kļūdu līmenis (WER - Word Error Rate) ir līdzīgs CER, bet tas tiek mērīts vārdu līmenī. WER aprēķina minimālo operāciju skaitu (ieskaitot ievietošanu, dzēšanu un aizstāšanu), kas nepieciešams, lai pārveidotu OCR rezultātu vārdos par oriģinālo tekstu. Tāpat kā CER, WER tiek aprēķināts, dalot šo operāciju skaitu ar oriģinālā teksta vārdu skaitu, un rezultāts tiek izteikts procentos.

CER un WER ir ļoti izplatītas kvantitatīvas OCR validācijas metodes, jo tās sniedz skaidru un tiešu mērījumu OCR precizitātei. Tomēr ir svarīgi atzīmēt, ka šīs metodes var būt mazāk efektīvas, ja OCR sistēma darbojas ar sarežģītu teksta izkārtojumu vai ja oriģinālajā tekstā ir struktūras (piemēram, tabulas vai attēli), kuras nav vienkārši pārveidojamas par lineāru teksta virkni. Šajos gadījumos var būt noderīgas papildu validācijas metodes.

3.3. Teksta salasāmība

Salasāmība ir viens no svarīgākajiem kvalitatīvajiem rādītājiem, lai novērtētu OCR rezultātu kvalitāti. Šī metrika mēra, cik viegli lasāms ir OCR izvadītais teksts, ņemot vērā ne tikai individuālo rakstzīmju un vārdu precizitāti, bet arī teksta struktūru un loģiku kopumā.

Salasāmība neaprobežojas tikai ar pareizu rakstzīmju atpazīšanu. Tā ietver arī citus faktorus, piemēram:

- teksta struktūra: teksts jāsaglabā pareizā formātā, tostarp paragrāfi, punktuācija, atstarpes starp vārdiem un citi formatēšanas elementi. OCR sistēma, kas pārveido attēlus par tekstu bez pareizas formatēšanas, var radīt tekstu, kas ir grūti lasīt un saprast;
- loģiska kohēzija: tekstam jābūt loģiski secīgam, lai būtu viegli lasāms. Ja OCR sistēma nepareizi atpazīst vārdus vai frāzes, tas var traucēt teksta nozīmi un padarīt to grūtāk saprotamu;
- pareiza gramatika un interpunkcija: pareiza gramatika un interpunkcija ir būtiska lasīšanas vieglumam. OCR sistēmas, kas nepareizi atpazīst punktuācijas zīmes vai neņem vērā gramatikas noteikumus, var radīt tekstu, kuru ir grūti lasīt;
- saglabāšana un atpazīšana: salasāmības mērījums var ietvert arī teksta saglabāšanu un atpazīšanu, piemēram, fonta izmēru un stilu, kā arī attēlu un diagrammu pareizu atpazīšanu un atveidošanu.

Salasāmības mērījums prasa subjektīvu novērtējumu, un tas var atšķirties atkarībā no konkrētās lietojumprogrammas vai lietotāja vajadzībām. Tomēr tas ir svarīgs rādītājs, lai novērtētu, cik efektīvi OCR sistēma var pārvērst attēlu par tekstu, kas ir viegli lasāms un saprotams.

4. TEKSTA APGABALU KLASIFICĒŠANA

Klasificēt teksta apgabalus ir nozīmīgi, analizējot latviešu etimoloģijas vārdnīcu, jo tās pamata alfabēts ir latviešu alfabēts, savukārt, lai risinātu svešvalodu alfabēta problēmas OCR pēcapstrādes posmā ir nozīmīgi klasificēt vārdus kuri nav pamata valodā (latviešu valodā). Svešvalodu alfabētu vārdus var atpazīt pēc sintaktiskajām teksta pazīmēm, kā arī pārzinot vārdnīcas struktūru.

K. Karuļa *Latviešu etimoloģijas vārdnīcas* [2] pamatteksts ir stingri strukturēts un aprakstīts 11. lpp. nodaļā: *VĀRDNĪCAS IEKĀRTOJUMS*. Šajā nodaļā tiek noteikts:

- šķirkļa sākumā ir analizējamais vārds un tā izruna (kvadrātiekavās);
- tālāk seko radu valodu vienas cilmes vārdi, kur vārda valodas apzīmējuma saīsinājums seko pirms paša vārda;
- pēc tam etimoloģiskā analīze mantotajiem vārdiem kura sākas ar norādi uz rekonstruēto indoeiropiešu pirmvalodas sakni un tās fonētiskajām pārmaiņām vēsturiskajā attīstībā;
- pēc vārda formas analīzes dota nozīmju attīstības shēma. Dažos gadījumos tā ilustrēta ar folkloras, seno rakstu vai izlokšņu piemēriem;
- šķirkļa beigās dotas literatūras norādes.

4.1. Teksta apgabalu definēšana

Izstrādājot darbu, autors veica svešvalodu alfabētu apgabalu klasifikāciju *Latviešu etimoloģijas vārdnīcā*. Šī klasifikācija sniedz strukturētu un sistematizētu informāciju par etimoloģijas vārdnīcā izmantotajiem svešvalodu alfabētu vārdiem, kas ir iekļauti vārdnīcā.

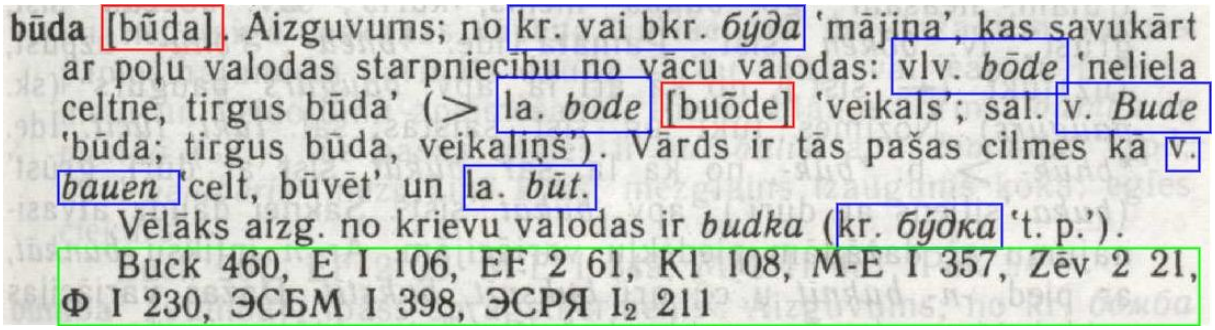
Svešvalodu alfabētu apgabali ir klasificēti vairākos veidos, atkarībā no to atrašanās vietas šķirkļa struktūrā un mērķa:

- šķirkļa sākumā, šķirkļa izrunas apgabalā tiek sniegta informācija par vārda izrunu;
- šķirkļa paskaidrojošajā pamattekstā, kur svešvalodā rakstītam vārdam tiek lietots svešvalodas apzīmējuma prefikss;
- šķirkļa paskaidrojošajā pamattekstā, kur tiek lietotas vārda izrunas formas vārda analīzei;
- šķirkļa pamatteksta pēdējā rindkopā, kur tiek dotas literatūras atsauces.

4.1.1. Pamatteksta apgabals

Pamatteksta apgabals ir vārdnīcas galvenais teksta apgabals, kas satur visbūtiskāko informāciju par šķirkli un tā analīzi. Šis apgabals ir rakstīts vārdnīcas pamatvalodā, kas ir latviešu valoda, un tam ir centrālā nozīme vārdnīcas struktūrā un saturā.

Tomēr ir jāņem vērā, ka pamatteksta apgabala sastāvā var būt arī svešvalodu vai svešvalodu alfabētu apgabali (sk. att. 4.1.). Šie apakšapgabali tiek iekļauti, lai sniegtu informāciju par svešvalodām, izcelsmi, fonētiskajiem aspektiem vai vārdu attiecībām ar citām valodām. Šajos apakšapgabalos var būt izmantoti svešvalodu apzīmējumi, svešvalodas alfabēti vai citi svešvalodas rakstības elementi.


būda [būda]. Aizguvums; no kr. vai bkr. *бýда* 'māja', kas savukārt ar poļu valodas starpniecību no vācu valodas: vlv. *bōde* 'neliela celtne, tirgus būda' (> la. *bode* [buōde] 'veikals'; sal. v. *Bude* 'būda; tirgus būda, veikaliņš'). Vārds ir tās pašas cilmes kā v. *bauen* 'celt, būvēt' un la. *būt*.

Vēlāks aizg. no krievu valodas ir *budka* (kr. *бýдка* 't. p.').

Buck 460, E I 106, EF 2 61, KI 108, M-E I 357, Zēv 2 21, Ф I 230, ЭСБМ I 398, ЭСРЯ I_2 211

Att. 4.1 Karulis, Konstantīns, Latviešu etimoloģijas vārdnīca. Rīga: Avots, 2001. 155. lpp. LNB digitalizējums. Pamatteksta apgabals. Sarkanā krāsā: izrunas apgabali. Zilā krāsā: svešvalodu apgabali. Zaļā krāsā: literatūras atsauču apgabals.

4.1.2. Izrunas teksta apgabals

Izrunas teksta apgabals etimoloģijas vārdnīcā tiek lietots, lai sniegtu izrunas definīciju konkrētam vārdam vai vārda daļai. Tas ir īpaši svarīgs, lai palīdzētu lasītājiem izprast un pareizi izrunāt vārdus, jo dažreiz latviešu valodā vārdu izruna var nebūt acīmredzama tikai pēc rakstības.

Izrunai etimoloģijas vārdnīcā tiek izmantots internacionālais fonētiskais alfabēts [4], kas ļauj precīzi atspoguļot skaņu un fonētisko struktūru. Tas nodrošina vienotu un skaidru informāciju par to, kā konkrēts vārds vai vārda daļa tiek izrunāta.

Lai vizuāli atdalītu izrunas apgabalus, *Latviešu etimoloģijas vārdnīcā* izmanto kvadrātiekavas. Šīs kvadrātiekavas iezīmē izrunas teksta apgabalu un nodrošina skaidru norādi, kura daļa ir izrunas definīcija. Attēlā 4.1. ir redzams piemērs, kā šie izrunas apgabali tiek vizuāli atdalīti.

4.1.3. Svešvalodu teksta apgabals

Svešvalodas teksta apgabals etimoloģijas vārdnīcā tiek izmantots, lai veiktu šķirkļa analīzi svešvalodas vārdiem. Šie apgabali ir paredzēti, lai sniegtu informāciju par vārdu izcelsmi un no kuras svešvalodas tie ir aizgūti. Šāda informācija ir svarīga, lai izprastu vārdu vēsturisko kontekstu un saistības ar citām valodām.

Svešvalodas apgabali etimoloģijas vārdnīcā ir definēti pēc to valodas saīsinājuma, kas tiek norādīts pirms paša svešvalodas vārda. Attēlā 4.1. ir redzams piemērs, kā tiek atdalīti svešvalodas apgabali.

Lai iegūtu konkrētu valodas saīsinājumu svešvalodas apgabalam, ir nepieciešams apskatīt *Latviešu etimoloģijas vārdnīcas* 18. lappusi [2]. Šajā sadaļā tiek sniegta detalizēta informācija par svešvalodas saīsinājumiem, kas tiek izmantoti šķirkļa analīzē.

4.1.4. Atsauču teksta apgabals

Atsauču teksta apgabals etimoloģijas vārdnīcā tiek izmantots, lai nodrošinātu trasējamību un atsauces informāciju saistībā ar šķirkļa analīzi. Šis apgabals atrodas šķirkļa paskaidrojošā pamatteksta pēdējā rindkopā, un tā mērķis ir sniegt literatūras avotu informāciju, kas ir svarīga, lai veiktu vēlākus pētījumus vai precīzāk izprast vārda etimoloģiskās detaļas.

Lai iegūtu konkrētus literatūras avotus, ir nepieciešams skatīt *Latviešu etimoloģijas vārdnīcas* 21. lappusi [2]. Šajā sadaļā ir sniegta informācija par literatūras avotu saīsinājumiem, kas tiek izmantoti atsauču teksta apgabalā.

4.2.Tekstu apgabalu OCR implementācija

Tekstu apgabali OCR sistēmā tiek izmantoti pēcapstrādes posmā, izņemot pamatteksta apgabalu, lai veiktu precīzu valodas identifikāciju un tās turpmāku apstrādi. Šie apgabali ir būtiski, lai sistēma varētu veikt detalizētu svešvalodas analīzi un interpretāciju attiecībā uz iegūto tekstu. Pēc teksta iegūšanas no attēla vai dokumenta OCR sistēma identificē dažādus teksta apgabalus, kas ietver vārdus un rindkopas. Šī identifikācija tiek veikta, pielietojot algoritmus, kas spēj atpazīt un izdalīt atsevišķus apgabalus no visa teksta.

4.2.1. Pamatteksta modulis

Pamatteksta apgabals ir būtisks un sākotnējais apgabals, kas kalpo kā pamatvalodas bāze un nosaka OCR sistēmas darbības inicializāciju. Šis apgabals tiek apstrādāts un atpazīts, izmantojot vārdnīcas pamatvalodas konfigurāciju, kas šajā gadījumā ir latviešu valoda.

Pamatteksta apgabals ne tikai kalpo kā bāzes apgabals, kas inicializē OCR sistēmas darbību, bet arī sniedz pamatu tekstu apgabalu identifikācijai un turpmākai apstrādei. Tas veicina precīzāku un efektīvāku teksta atpazīšanu un ļauj sistēmai koncentrēties uz vajadzīgajiem teksta elementiem vai struktūrām, nodrošinot augstāku precizitāti un efektivitāti visā OCR procesā.

4.2.2. Izrunas modulis

Izrunas teksta apgabals OCR sistēmā tiek balstīts uz spēju atpazīt kvadrātiekavas, kas norāda uz vārda fonētisko skanējumu. Šis apgabals tiek iekļauts OCR pēcapstrādes posmā, kur katrs atpazītais vārds tiek pārbaudīts lineāri, un tiek meklētas kvadrātiekavas, kas atzīmē vārda izrunu. Kad šīs kvadrātiekavas ir atrastas, tiek veikta analīze, lai noteiktu šī vārda koordinātas attēlā, un pēc tam vārds tiek izgriezts no attēla.

Izgrieztajam vārdam tiek veikta tālāka apstrāde OCR pēcapstrādes posmā, kur tas tiek atpazīts visās pieejamajās OCR valodu konfigurācijās. Šajā posmā tiek izmantots *Tesseract* OCR rīks, kas pašlaik nespēj izmantot fonētiskā alfabēta modeli. Tāpēc vārdi tiek atpazīti, balstoties uz pieejamajiem valodu modeļiem [5].

Visi atpazīto vārdu valodu konfigurāciju rezultāti tiek apkopoti, un tiek izvēlēts tas rezultāts, kuram attiecīgā OCR sistēma piešķir augstāko pārliecības novērtējumu. Tas ļauj noteikt visuzticamāko vārda atpazīšanas rezultātu, kas ir balstīts uz dotās OCR sistēmas izvērtējumu.

4.2.3. Svešvalodu modulis

Svešvalodu teksta apgabalu identifikācijai tiek izmantota speciāli definēta vārdnīca, kas nodrošina saikni starp *Tesseract* OCR sistēmas pieejamajām valodām un vārdnīcā noteiktajiem valodu apzīmējumu saīsinājumiem (sk. 1. pielikumu). Šī vārdnīca ir būtiska OCR pēcapstrādes posmā, kur katram atpazītajam vārdam no pamatteksta apgabala tiek meklēts atbilstošs saīsinājums vārdnīcā.

Ja tiek atrasts vārds, kurš sakrīt ar vārdnīcas definēto saīsinājumu, sistēma identificē to kā svešvalodu teksta apgabalu. Šajā gadījumā tiek veikta attēla izgriešana, un šis izgrieztais vārds tiek atpazīts, izmantojot atbilstošo svešvalodas OCR konfigurāciju. Gala rezultātā tiek saglabāts tas vārds, kuru OCR sistēma novērtē ar augstāku pārliecības novērtējumu. Tas nozīmē, ka, ja tiek atpazīts vārds, kas atbilst vārdnīcā definētajam saīsinājumam un tas tiek atpazīts ar labāku uzticamību, šis vārds tiek saglabāts kā gala rezultāts.

Šāda metode, kur tiek izmantota vārdnīca un saīsinājumu saskaņošana, ļauj precīzāk identificēt svešvalodu teksta apgabalus un nodrošina izvēli starp atbilstošajām OCR valodu

konfigurācijām, lai panāktu vislabākos rezultātus. Tas nodrošina efektīvāku un precīzāku svešvalodu teksta atpazīšanu OCR sistēmā, ļaujot iegūt vērtīgu informāciju par valodas izcelsmi un īpašībām.

4.2.4. Atsauču modulis

Atsauču teksta apgabalu identifikācijai tiek izmantoti specifiski teksta izklājuma parametri, kas tiek iegūti no pamatteksta apgabala. Šie izklājuma parametri tiek analizēti, lai veiktu pamatteksta apgabala sadalīšanu rindkopās, kurās pēdējā rindkopa ir atsauču rindkopa. Tādējādi, pamatojoties uz formatējumu un struktūru, tiek noteikts, kur tieši atrodas atsauču teksta apgabals.

Kad ir identificēta pēdējā rindkopa kā atsauču teksta apgabals, katram vārdam šajā rindkopā tiek veikta izgriešana un tiek pielietota OCR sistēma, izmantojot visu pieejamo valodu konfigurāciju. Tas nozīmē, ka atsauču teksta vārdiem tiek veikta pilnīga atpazīšana, izmantojot visas pieejamās valodas un to specifiskos OCR parametrus.

Pēc tam tiek apkopoti OCR rezultāti, un no tiem tiek izvēlēti vārdi ar augstāko OCR pārliecības novērtējumu. Šie vārdi tiek aizvietoti atsauču apgabalā gala rezultātā, un pamatteksts tiek papildināts ar atsauču apgabala atjauninājumiem.

5. OCR PĒCAPSTRĀDES SISTĒMAS IZSTRĀDE

Sistēmas izstrādes pamatā tiek izmantots *Tesseract* OCR rīks, kas ir plaši izmantota atvērtā koda optiskās rakstības atpazīšanas tehnoloģija. Lai panāktu labākos rezultātus, *Tesseract* valodu konfigurācijās tiek izmantoti *tesseract_best* kategorijas valodu modeļi [5]. *Tesseract_best* valodu modeļi nodrošina augstu precizitāti un kvalitāti teksta atpazīšanā, īpaši attiecībā uz sarežģītiem tekstiem. Tie ir optimizēti, lai sniegtu labus rezultātus dažādu valodu atpazīšanā. Tomēr, ņemot vērā to augsto precizitāti, darbības ātrums ir lēnāks salīdzinājumā ar citiem *Tesseract* kategoriju valodu modeļiem.

Bakalaura darba ietvaros tiek veikta attēlu segmentācija. Šim nolūkam tika izmantots LNB *Latviešu etimoloģijas vārdnīcas* digitālais saturs [2], kas tika nodrošināts bakalaura darba izstrādes ietvaros. Šis saturs ietvēra attēlus no grāmatas lapaspušu skenējumiem un *alto* formāta struktūrfailus.

Programmatūra, kas tiek izstrādāta bakalaura darba ietvaros, balstās uz *Python* (v3.10.6) programmēšanas valodu. *Python* ir populāra un plaši izmantota valoda, kas nodrošina daudzpusību, efektivitāti un plašu bibliotēku atbalstu. Lai integrētu *Tesseract* rīku programmatūrā un nodrošinātu tā izmantošanu, tiek izmantota *pytesseract* (v0.3.10) pakotne. *Pytesseract* ir *Python* bibliotēka, kas sniedz iespēju saziņai ar *Tesseract* OCR rīku un veic teksta atpazīšanu attēlos. Tā nodrošina ērtu un vienkāršu veidu, kā izmantot *Tesseract* funkcionalitāti *Python* vidē.

5.1. Pamatteksta modulis

Pamatteksta modulis, kas ir svarīga programmas sastāvdaļa, sastāv no divām galvenajām daļām: attēla segmentācijas un OCR rīka inicializācijas. Šīs daļas ir atbildīgas par attēla apstrādi un teksta atpazīšanu.

Attēla segmentācijas posmā tiek izmantots LNB sniegtais *alto* struktūrfails [6], kas satur informāciju par lapaspuses izkārtojumu un teksta novietojumu attēlā. No šī struktūrfaila tiek ņemti *PrintSpace* taga atribūti, kas norāda uz teksta lauka novietojumu attēlā. Šie atribūti tiek izmantoti, lai iegūtu koordinātas, kas definē katru lapaspuses attēla segmentu.

PrintSpace tags ir arī izmantots LNB digitālajā bibliotēkā kā pazīme (sk. att. 5.1.), pēc kuras OCR rīks atpazīst teksta laukus attēlā. Tas ir svarīgi, lai noteiktu, kurās vietās attēlā atrodas teksts, ko OCR rīks varēs pareizi atpazīt. Šī informācijas iegūšana no *alto* struktūrfaila ļauj precīzi segmentēt attēlu un koncentrēties uz būtisko teksta informāciju.

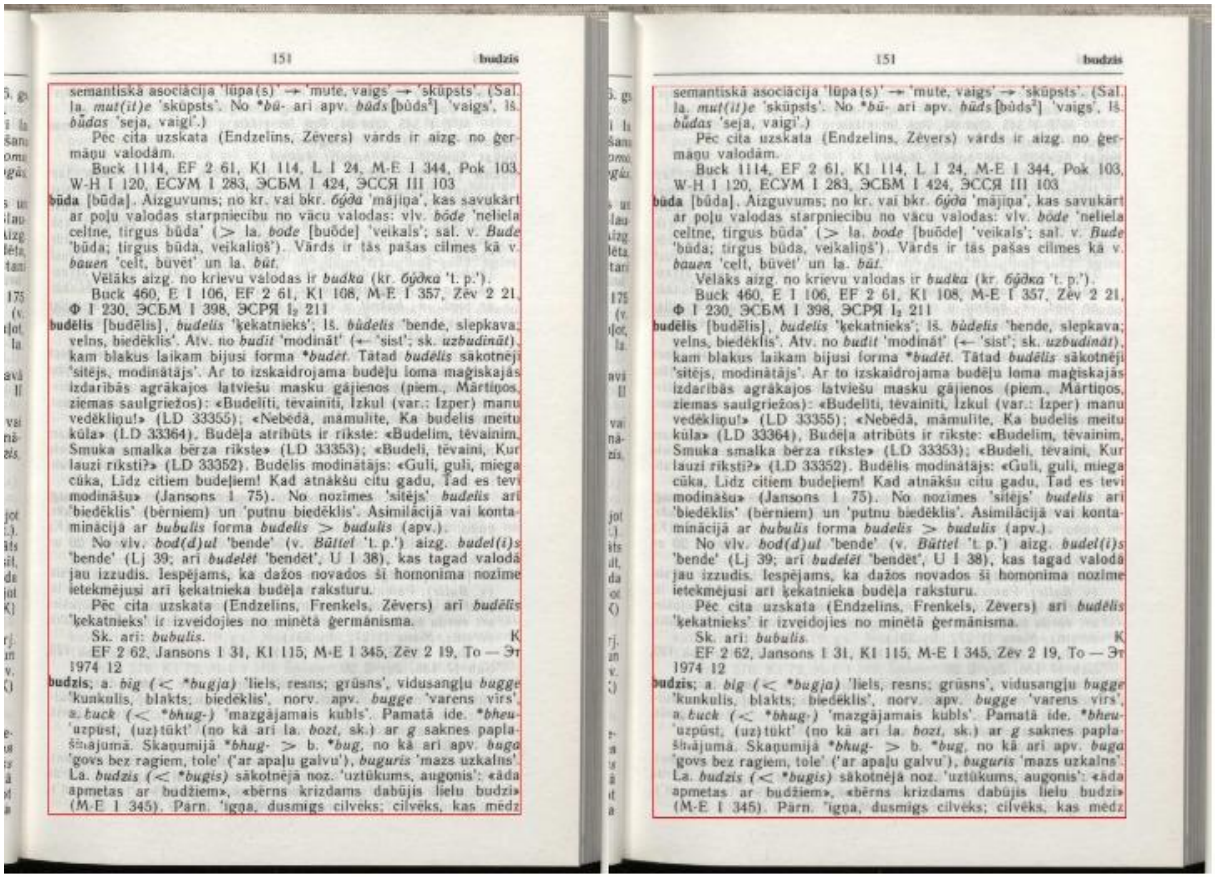
Kad attēls ir veiksmīgi segmentēts, tas tiek nodots *pytesseract* pakotnes *image_to_data* metodei, kas ļauj veikt teksta atpazīšanu attēlā. Šī metode atgriež datu kopu ar atpazītajiem

vārdiem, to novietojumu attēlā, pārliecības novērtējumu un citiem noderīgiem parametriem (sk. tab. 5.1.). Šie dati ir svarīgi turpmākai pēcapstrādei un analīzei.

Tab. 5.1. dataFrame objekta izgriezums pēc image_to_data metodes izpildes.

	level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
4	5	1	1	1	1	1	108	14	227	57	90	pastalas
5	5	1	1	1	1	2	388	14	154	55	85	žaļkst
6	5	1	1	1	1	3	597	27	58	30	96	un
7	5	1	1	1	1	4	705	14	156	50	92	vēlāk,
8	5	1	1	1	1	5	915	15	307	55	88	pārliecigam

Pēc datu kopas iegūšanas no pytesseract metodes, tiek uzsākts pēcapstrādes process. Tas var ietvert rezultātu apstrādi, korekcijas vai citus pielāgojumus, lai iegūtu optimālus rezultātus. Pēcapstrādes posms ir būtisks, lai nodrošinātu precīzu un kvalitatīvu teksta atpazīšanu un sniegtu labākos rezultātus programmā.

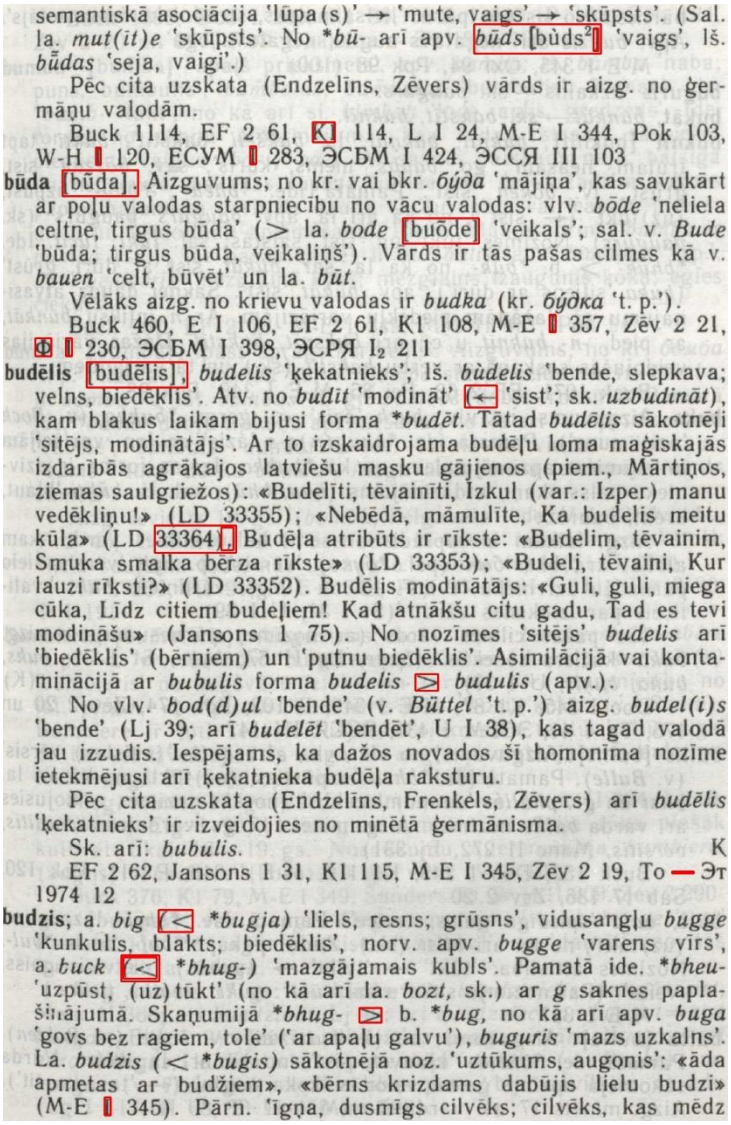


Att. 5.1. Karulis, Konstantīns, Latviešu etimoloģijas vārdnīca. Rīga: Avots, 2001. 151. lpp. LNB digitalizējums. Labajā pusē LNB digitālās bibliotēkas digitalizētais attēla apgabals. Kreisajā pusē pamatteksta moduļa PrintSpace apgabals.

5.2.Izrunas modulis

Izrunas modulis tiek aktivizēts OCR pēcapstrādes procesa laikā, izmantojot pamatteksta moduļa iegūto datu kopu (sk. tabulu 5.1.). Šis modulis ir atbildīgs par izrunas teksta apgabala identifikāciju, izmantojot kvadrātiekavu meklēšanu, kas atrodas *text* kolonnā. Modulis meklē kvadrātiekavas datu kopas rindās, un ja tās tiek atrastas, tas nozīmē, ka tajā vietā atrodas vārds, kam ir izrunas definīcija (sk. att. 5.2.)

Ja kvadrātiekava tiek atrasta, tiek meklēta otra (aizverošā vai atverošā) kvadrātiekava vienas datu kopas rindas intervālā. Ja otra kvadrātiekava tiek atrasta, tad izrunas modulis apvieno divas rindas (atverošo un aizverošo kvadrātiekavu rindas) vienā rindā, lai saglabātu datu integritāti. Tas ļauj apkopot abu rindu informāciju par vārda izrunu un turpināt vārda turpmāku apstrādi.



semantiskā asociācija 'lūpa(s)' → 'mute, vaigs' → 'skūpsts'. (Sal. la. *mut(īt)e* 'skūpsts'. No **bū-* arī apv. *būds* [bùds²] 'vaigs', lš. *bũdas* 'seja, vaigi'.)

Pēc cita uzskata (Endzelīns, Zēvers) vārds ir aizg. no ģermāņu valodām.

Buck 1114, EF 2 61, K 114, L I 24, M-E I 344, Pok 103, W-H I 120, ЕСУМ I 283, ЭСБМ I 424, ЭССЯ III 103

būda [būda], Aizguvums; no kr. vai bkr. *бýда* 'mājiņa', kas savukārt ar poļu valodas starpniecību no vācu valodas: vlv. *bōde* 'neliela celtne, tirgus būda' (> la. *bode* [buōde] 'veikals'; sal. v. *Bude* 'būda; tirgus būda, veikaliņš'). Vārds ir tās pašas cilmes kā v. *bauen* 'celt, būvēt' un la. *būt*.

Vēlāks aizg. no krievu valodas ir *budka* (kr. *бýдка* 't. p.').

Buck 460, E I 106, EF 2 61, Kl 108, M-E I 357, Zēv 2 21, Ф I 230, ЭСБМ I 398, ЭСРЯ I_2 211

budēlis [budēlis], *budelis* 'ķekatnieks'; lš. *bùdelis* 'bende, slepkava; velns, biedēklis'. Atv. no *budit* 'modināt' (← 'sist'; sk. *uzbudināt*), kam blakus laikam bijusi forma **budēt*. Tātad *budēlis* sākotnēji 'sitējs, modinātājs'. Ar to izskaidrojama budēļu loma maģiskajās izdarībās agrākajos latviešu masku gājienos (piem., Mārtiņos, ziemas saulgriežos): «Budelīti, tēvainīti, Izkul (var.: Izper) manu vedēkliņu!» (LD 33355); «Nebēdā, māmulīte, Ka budelis meitu kūla» (LD 33364). Budēļa atribūts ir rīkste: «Budelim, tēvainim, Smuka smalka bērza rīkste» (LD 33353); «Budeli, tēvaini, Kur lauzi rīksti?» (LD 33352). Budēlis modinātājs: «Guli, guli, miega cūka, Līdz citiem budeļiem! Kad atnākšu citu gadu, Tad es tevi modināšu» (Jansons I 75). No nozīmes 'sitējs' *budelis* arī 'biedēklis' (bērniem) un 'putnu biedēklis'. Asimilācijā vai kontaminācijā ar *bubulis* forma *budelis* ⊃ *budulis* (apv.).

No vlv. *bod(d)ul* 'bende' (v. *Büttel* 't. p.') aizg. *budel(i)s* 'bende' (Lj 39; arī *budelēt* 'bendēt', U I 38), kas tagad valodā jau izzudis. Iespējams, ka dažos novados šī homonīma nozīme ietekmējusi arī ķekatnieka budēļa raksturu.

Pēc cita uzskata (Endzelīns, Frenkels, Zēvers) arī *budēlis* 'ķekatnieks' ir izveidojies no minētā ģermānisma.

Sk. arī: *bubulis*. K

EF 2 62, Jansons I 31, Kl 115, M-E I 345, Zēv 2 19, То — Эт 1974 12

budzis; a. *big* (< **bugja*) 'liels, resns; grūsns', vidusangļu *bugge* 'kunkulis, blakts; biedēklis', norv. apv. *bugge* 'varens vīrs', a. *buck* (< **bhug-*) 'mazgājamais kubls'. Pamatā ide. **bheu-* 'uzpūst, (uz)tūkt' (no kā arī la. *bozt*, sk.) ar *g* saknes paplašinājumā. Skaņumijā **bhug-* ⊃ b. **bug*, no kā arī apv. *buga* 'govs bez ragiem, tole' ('ar apaļu galvu'), *buguris* 'mazs uzkalns'. La. *budzis* (< **bugis*) sākotnējā noz. 'uztūkums, augonis': «āda apmetas ar budžiem», «bērns krizdams dabūjis lielu budzi» (M-E I 345). Pārn. 'īgņa, dusmīgs cilvēks; cilvēks, kas mēdz

Att. 5.2. Karulis, Konstantīns, Latviešu etimoloģijas vārdnīca. Rīga: Avots, 2001. 155. lpp. LNB digitalizējums. Sarkanā krāsā izrunas teksta apgabali, kas apstrādāti izrunas modulī.

Nākamajā solī tiek veikta vārda izgriešana no attēla un tālāka apstrāde, pamatojoties uz svešvalodas modulim izveidotu valodu vārdnīcu, kura satur *tesseract_best* valodu modeļu apzīmējumus un vārdnīcas valodu atslēgvārdus. Vārda attēls tiek padots Tesseract OCR rīkam un tiek veiktas vairākas atpazīšanas reizes, mainot OCR valodas konfigurāciju katrā reizē atbilstoši katrai valodu vārdnīcas valodai.

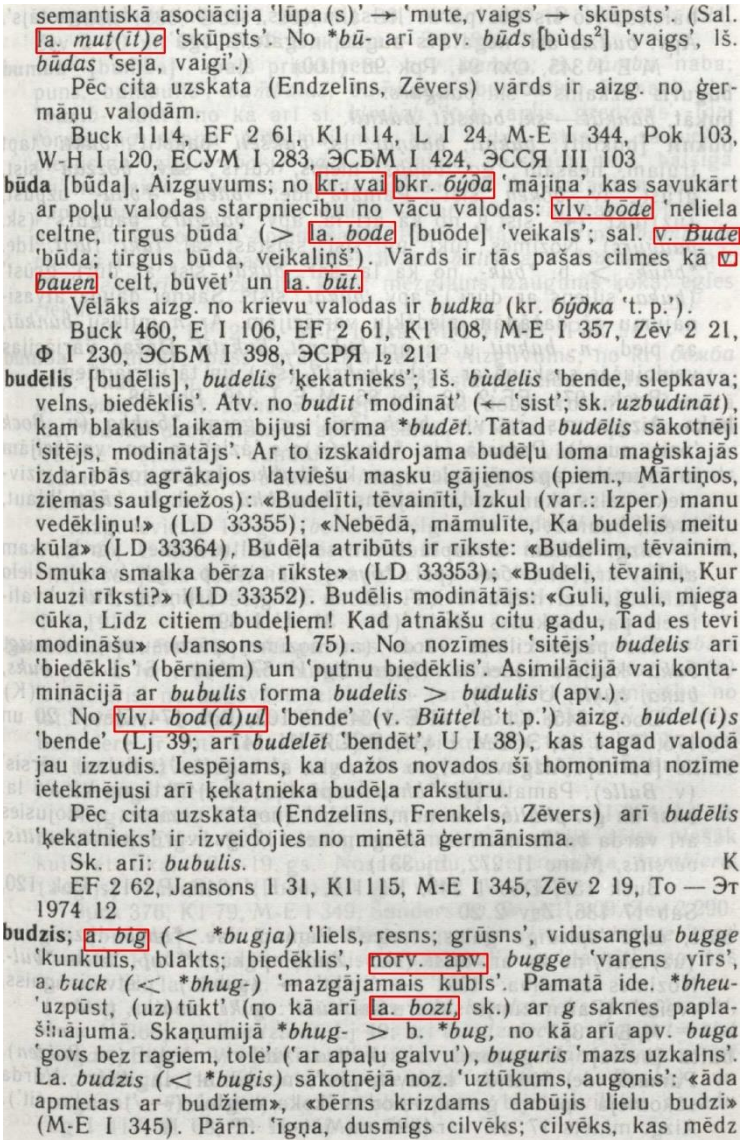
Pēc vārda apstrādes rezultātu iegūšanas, tiek veikta to apkopošana, un vārds ar augstāko pārliecības novērtējumu no OCR rīka tiek ievietots atpakaļ kopējā datu kopā. Tas nodrošina, ka galīgais rezultāts iekļauj vārdu ar visu iepriekšējo apstrādes posmu informāciju un augstu pārliecības novērtējumu no OCR rīka.

5.3.Svešvalodu modulis

Svešvalodu modulis ir svarīgs OCR pēcapstrādes procesa posms. Tas izmanto pamatteksta moduļa iegūto datu kopu (skat. tabulu 5.1.), lai identificētu svešvalodu apgabalus. Šajā modulī tiek izmantota valodu bibliotēka, kurā ir definēti valodu saīsinājumi no etimoloģijas vārdnīcas. Katram datu kopas ierakstam teksta kolonnā tiek salīdzināti šie valodu saīsinājumi.

Ja kāds ieraksts sakrīt ar kādu no valodu saīsinājumiem, algoritms meklē vārdu, kas seko šim saīsinājumam (sk. att. 5.3.). Tad no attēla tiek izgriezts attiecīgā vārda attēls, un tiek veikta atpazīšana, izmantojot atbilstošo svešvalodas OCR konfigurāciju. Ja vārds tiek atpazīts ar augstāku pārliecības novērtējumu nekā pamata valodā, datu kopa tiek atjaunināta ar jauniegūtajām vērtībām.

Šī procesa rezultātā datu kopa tiek pilnveidota, iekļaujot precīzāku svešvalodu informāciju, kura tika iegūta, izmantojot atpazīšanu ar augstu pārliecības novērtējumu no svešvalodas OCR rīka. Tas nodrošina, ka gala rezultāts ietver visaptverošu un precīzu informāciju par vārdiem arī svešvalodās.



semantiskā asociācija 'lūpa(s)' → 'mute, vaigs' → 'skūpsts'. (Sal. la. *mut(it)e* 'skūpsts'. No **bū-* arī apv. *būds* [bùds²] 'vaigs', lš. *būdas* 'seja, vaigi'.)

Pēc cita uzskata (Endzelīns, Zēvers) vārds ir aizg. no ģermāņu valodām.

Buck 1114, EF 2 61, Kl 114, L I 24, M-E I 344, Pok 103, W-H I 120, ЕСУМ I 283, ЭСБМ I 424, ЭССЯ III 103

būda [būda]. Aizguvums; no kr. vai bkr. *бýда* 'mājiņa', kas savukārt ar poļu valodas starpniecību no vācu valodas: vlv. *bōde* 'neliela celtne, tirgus būda' (> la. *bode* [buōde] 'veikals'; sal. v. *Bude* 'būda; tirgus būda, veikaliņš'). Vārds ir tās pašas cilmes kā v. *bauen* 'celt, būvēt' un la. *būt*.

Vēlāks aizg. no krievu valodas ir *budka* (kr. *бýдка* 't. p.').

Buck 460, E I 106, EF 2 61, Kl 108, M-E I 357, Zēv 2 21, Ф I 230, ЭСБМ I 398, ЭСРЯ I₂ 211

budēlis [budēlis], *budelis* 'ķekatnieks'; lš. *bùdelis* 'bende, slepkava; velns, biedēklis'. Atv. no *budit* 'modināt' (← 'sist'; sk. *uzbudināt*), kam blakus laikam bijusi forma **budēt*. Tātad *budēlis* sākotnēji 'sitējs, modinātājs'. Ar to izskaidrojama budēļu loma maģiskajās izdarībās agrākajos latviešu masku gājienos (piem., Mārtiņos, ziemas saulgriežos): «Budelīti, tēvainīti, Izkul (var.: Izper) manu vedēkliņu!» (LD 33355); «Nebēdā, māmulīte, Ka budelis meitu kūla» (LD 33364). Budēļa atribūts ir rīkste: «Budelim, tēvainim, Smuka smalka bērza rīkste» (LD 33353); «Budeli, tēvaini, Kur lauzi rīksti?» (LD 33352). Budēlis modinātājs: «Guli, guli, miega cūka, Līdz citiem budeļiem! Kad atnākšu citu gadu, Tad es tevi modināšu» (Jansons I 75). No nozīmes 'sitējs' *budelis* arī 'biedēklis' (bērniem) un 'putnu biedēklis'. Asimilācijā vai kontaminācijā ar *bubulis* forma *budelis* > *budulis* (apv.).

No vlv. *bod(d)ul* 'bende' (v. *Büttel* 't. p.') aizg. *budel(i)s* 'bende' (Lj 39; arī *budelēt* 'bendēt', U I 38), kas tagad valodā jau izzudis. Iespējams, ka dažos novados šī homonīma nozīme ietekmējusi arī ķekatnieka budēļa raksturu.

Pēc cita uzskata (Endzelīns, Frenkels, Zēvers) arī *budēlis* 'ķekatnieks' ir izveidojies no minētā ģermānisma.

Sk. arī: *bubulis*. K

EF 2 62, Jansons I 31, Kl 115, M-E I 345, Zēv 2 19, То — Эт 1974 12

budzis; a. *big* (< **bugja*) 'liels, resns; grūsns', vidusangļu *bugge* 'kunkulis, blakts; biedēklis', norv. apv. *bugge* 'varens vīrs', a. *buck* (< **bhug-*) 'mazgājamais kubls'. Pamatā ide. **bheu-* 'uzpūst, (uz)tūkt' (no kā arī la. *bozt*, sk.) ar *g* saknes paplašinājumā. Skaņumijā **bhug-* > b. **bug*, no kā arī apv. *buga* 'govs bez ragiem, tole' ('ar apaļu galvu'), *buguris* 'mazs uzkalns'. La. *budzis* (< **bugis*) sākotnējā noz. 'uztūkums, augonis': «āda apmetas ar budžiem», «bērns krizdams dabūjis lielu budzi» (M-E I 345). Pārn. 'īgņa, dusmīgs cilvēks; cilvēks, kas mēdz

Att. 5.3. Karulis, Konstantīns, Latviešu etimoloģijas vārdnīca. Rīga: Avots, 2001. 155. lpp. LNB digitalizējums. Sarkanā krāsā svešvalodu teksta apgabali, kas apstrādāti svešvalodu modulī.

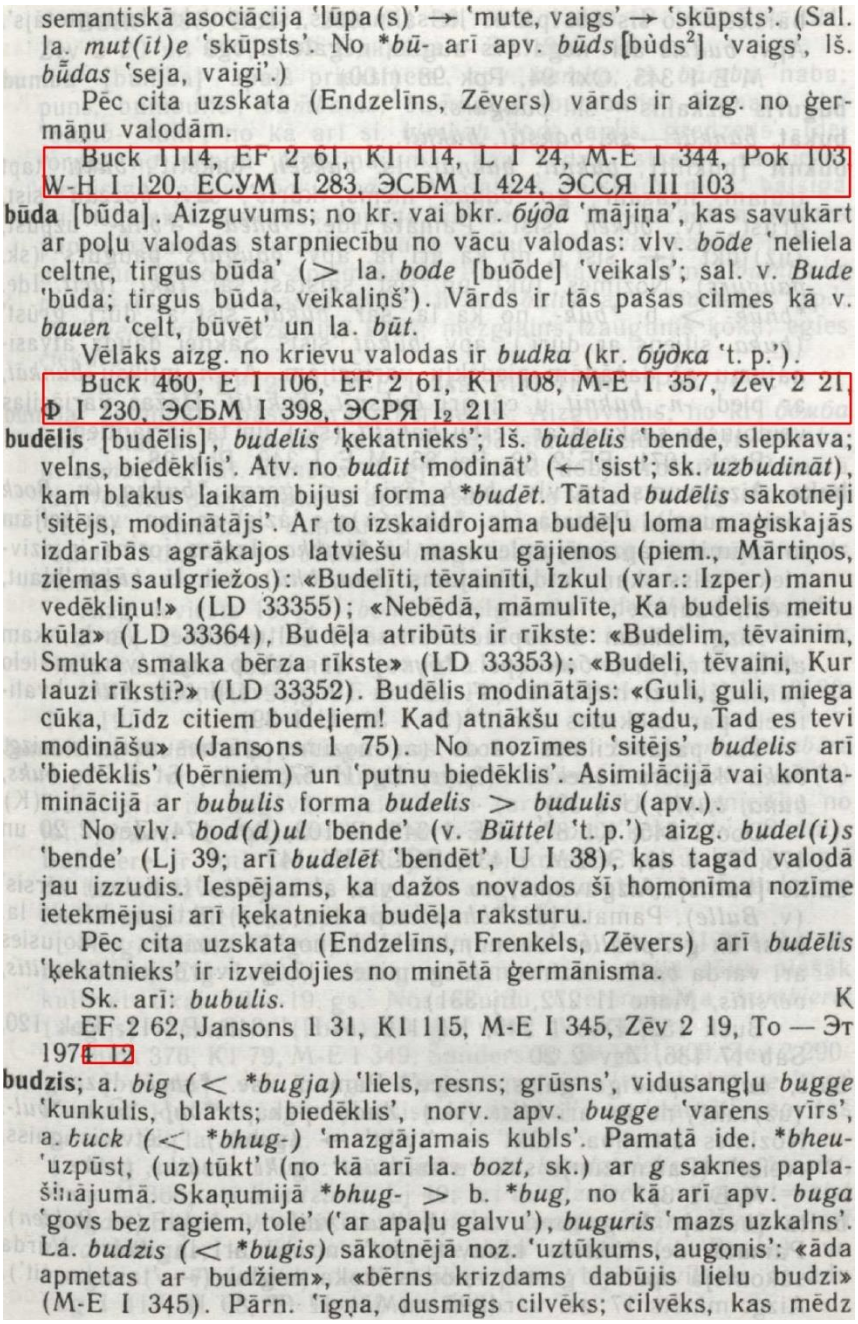
5.4. Atsauču modulis

Atsauču modulis ir viens no OCR pēcapstrādes moduļiem, kas veic analīzi pamatteksta moduļa iegūtajai datu kopai, ņemot vērā rindkopu izkārtojumu. Šī moduļa mērķis ir identificēt šķirkļa pēdējo rindkopu, izmantojot sekojošu algoritmu: vispirms pamatteksts tiek sadalīts šķirkļu pamattekstos, kur šķirkļa sākums ir manāmi izvirzīts uz kreiso pusi. Pēc tam šķirkļa pamatteksts tiek sadalīts rindkopās, un no tām tiek atlasīta pēdējā rindkopa, kas tiek noteikta kā atsauču teksta apgabals.

Pēc atsauču teksta apgabala identificēšanas (sk. att. 5.4.) katrs atlasītais vārds tiek izgriezts no attēla un apstrādāts, izmantojot visas pieejamās OCR valodu konfigurācijas. Pēc tam tiek meklēta tāda valodas konfigurācija, kas nodrošina lielāko pārliecības novērtējumu no

OCR rīka. Beigās vārds ar visaugstāko pārliecības novērtējumu tiek ievietots kopējā datu kopā, kas saglabā informāciju par visiem atpazītajiem vārdiem.

Šī moduļa darbība nodrošina pareizāku un uzticamāku atsauču teksta apgabalu, kur tiek izmantots OCR rīka augstākais pārliecības novērtējums, lai iegūtu visprecīzākos rezultātus. Tas palīdz uzlabot datu kopu, iekļaujot tikai vārdus ar vislielāko uzticamību no OCR atpazīšanas procesa.



semantiskā asociācija 'lūpa(s)' → 'mute, vaigs' → 'skūpsts'. (Sal. la. *mut(īt)e* 'skūpsts'. No **bū-* arī apv. *būds* [bùds[2]] 'vaigs', lš. *bũdas* 'seja, vaigi'.)

Pēc cita uzskata (Endzelīns, Zēvers) vārds ir aizg. no ģermāņu valodām.

Buck 1114, EF 2 61, KI 114, L I 24, M-E I 344, Pok 103, W-H I 120, ЕСУМ I 283, ЭСБМ I 424, ЭССЯ III 103

būda [būda]. Aizguvums; no kr. vai bkr. *бýда* 'mājiņa', kas savukārt ar poļu valodas starpniecību no vācu valodas: vlv. *bōde* 'neliela celtne, tirgus būda' (> la. *bode* [buōde] 'veikals'; sal. v. *Bude* 'būda; tirgus būda, veikaliņš'). Vārds ir tās pašas cilmes kā v. *bauen* 'celt, būvēt' un la. *būt*.

Vēlāks aizg. no krievu valodas ir *budka* (kr. *бýдка* 't. p.').

Buck 460, E I 106, EF 2 61, KI 108, M-E I 357, Zēv 2 21, Ф I 230, ЭСБМ I 398, ЭСРЯ I_2 211

budēlis [budēlis], *budelis* 'ķekatnieks'; lš. *bùdelis* 'bende, slepkava; velns, biedēklis'. Atv. no *budīt* 'modināt' (← 'sist'; sk. *uzbudināt*), kam blakus laikam bijusi forma **budēt*. Tātad *budēlis* sākotnēji 'sitējs, modinātājs'. Ar to izskaidrojama budēļu loma maģiskajās izdarībās agrākajos latviešu masku gājienos (piem., Mārtiņos, ziemas saulgriežos): «Budelīti, tēvainīti, Izkul (var.: Izper) manu vedēkliņu!» (LD 33355); «Nebēdā, māmulīte, Ka budelis meitu kūla» (LD 33364). Budēļa atribūts ir rīkste: «Budelim, tēvainim, Smuka smalka bērza rīkste» (LD 33353); «Budeli, tēvaini, Kur lauzi rīksti?» (LD 33352). Budēlis modinātājs: «Guli, guli, miega cūka, Līdz citiem budeļiem! Kad atnākšu citu gadu, Tad es tevi modināšu» (Jansons I 75). No nozīmes 'sitējs' *budelis* arī 'biedēklis' (bērniem) un 'putnu biedēklis'. Asimilācijā vai kontaminācijā ar *bubulis* forma *budelis* > *budulis* (apv.).

No vlv. *bod(d)ul* 'bende' (v. *Büttel* 't. p.') aizg. *budel(i)s* 'bende' (Lj 39; arī *budelēt* 'bendēt', U I 38), kas tagad valodā jau izzudis. Iespējams, ka dažos novados šī homonīma nozīme ietekmējusi arī ķekatnieka budēļa raksturu.

Pēc cita uzskata (Endzelīns, Frenkels, Zēvers) arī *budēlis* 'ķekatnieks' ir izveidojies no minētā ģermānisma.

Sk. arī: *bubulis*. K

EF 2 62, Jansons I 31, KI 115, M-E I 345, Zēv 2 19, То — Эт 1974 12

budzis; a. *big* (< **bugja*) 'liels, resns; grūsns', vidusangļu *bugge* 'kunkulis, blakts; biedēklis', norv. apv. *bugge* 'varens vīrs', a. *buck* (< **bhug-*) 'mazgājamais kubls'. Pamatā ide. **bheu-* 'uzpūst, (uz)tūkt' (no kā arī la. *bozt*, sk.) ar *g* saknes paplašinājumā. Skaņumijā **bhug-* > b. **bug*, no kā arī apv. *buga* 'govs bez ragiem, tole' ('ar apaļu galvu'), *buguris* 'mazs uzkalns'. La. *budzis* (< **bugis*) sākotnējā noz. 'uztūkums, augonis': «āda apmetas ar budžiem», «bērns krizdams dabūjis lielu budzi» (M-E I 345). Pārn. 'īgņa, dusmīgs cilvēks; cilvēks, kas mēdz

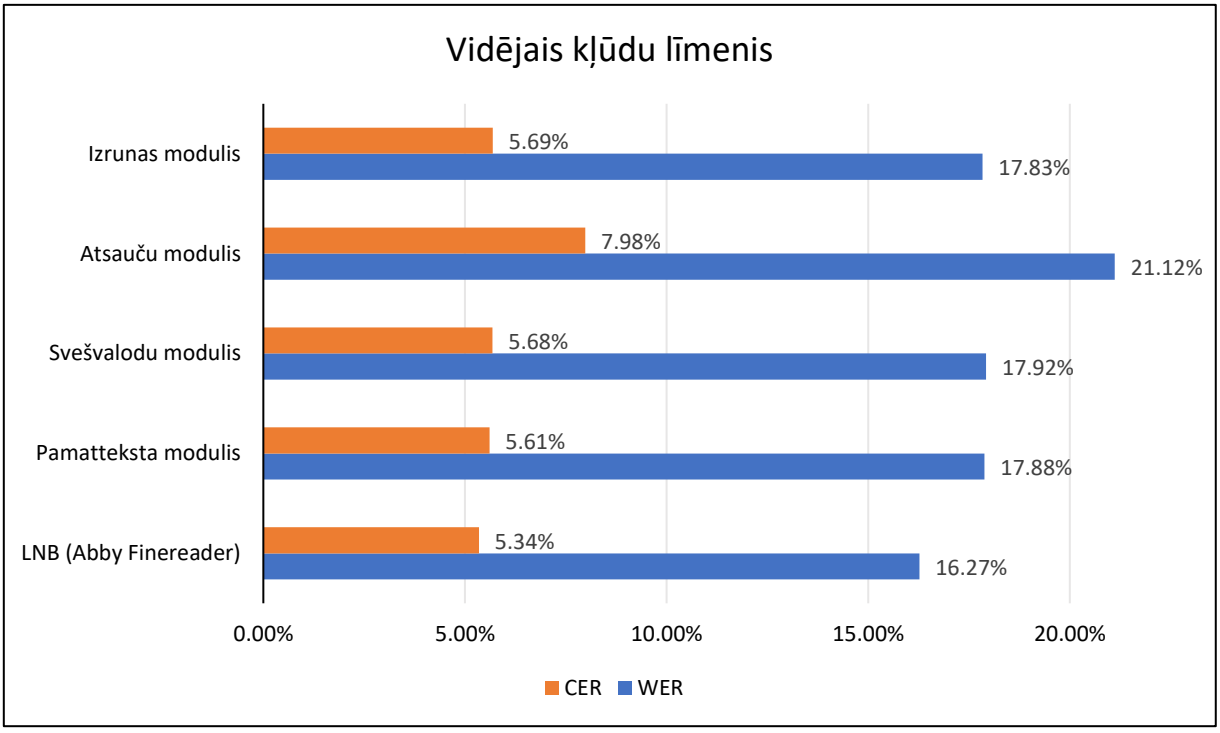
Att. 5.4. Karulis, Konstantīns, Latviešu etimoloģijas vārdnīca. Rīga: Avots, 2001. 155. lpp. LNB digitalizējums. Sarkanā krāsā atsauču teksta apgabali, kas apstrādāti atsauču modulī.

REZULTĀTI

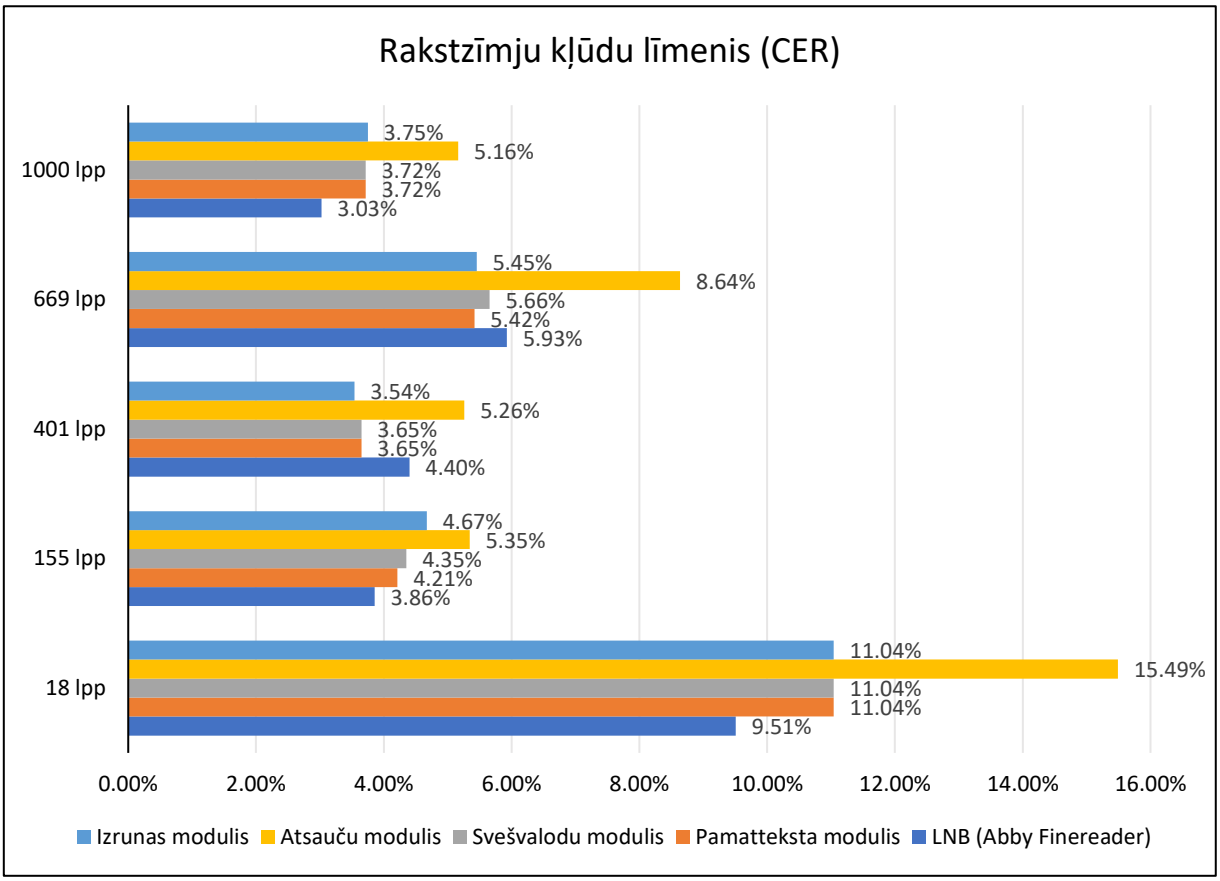
Lai pārbaudītu OCR pēcapstrādes sistēmas rezultātus, tika veikta manuāla datu ievade piecām lapaspusēm no *Latviešu etimoloģijas vārdnīcas*. Šīs lapaspuses tika izvēlētas pēc nejaušības principa, nodrošinot reprezentatīvu izlasi. Veicot manuālo datu ievadi, tika veikta rūpīga rakstzīmju transkripcija, lai panāktu maksimālu līdzību starp ievadītajām rakstzīmēm un attēlā attēlotajām rakstzīmēm. Šāda precizitāte nodrošina ticamus un salīdzināmus datu rezultātus, kas ļauj objektīvi novērtēt OCR pēcapstrādes sistēmas veiktos rezultātus (sk. 1. pielikumu).

Kvantitatīvo metriku rezultātu iegūšanai tika izveidots speciāls rezultātu validācijas modulis, kas nodrošina objektīvu OCR rīka rezultātu pārbaudi, salīdzinot to ar manuāli ievadītajām lapaspusēm. Šis modulis veic detalizētu analīzi, izmantojot salīdzināšanas algoritmus, lai noteiktu precizitāti, atbilstību un līdzību starp OCR rezultātiem un manuāli ievadītajām lapaspusēm.

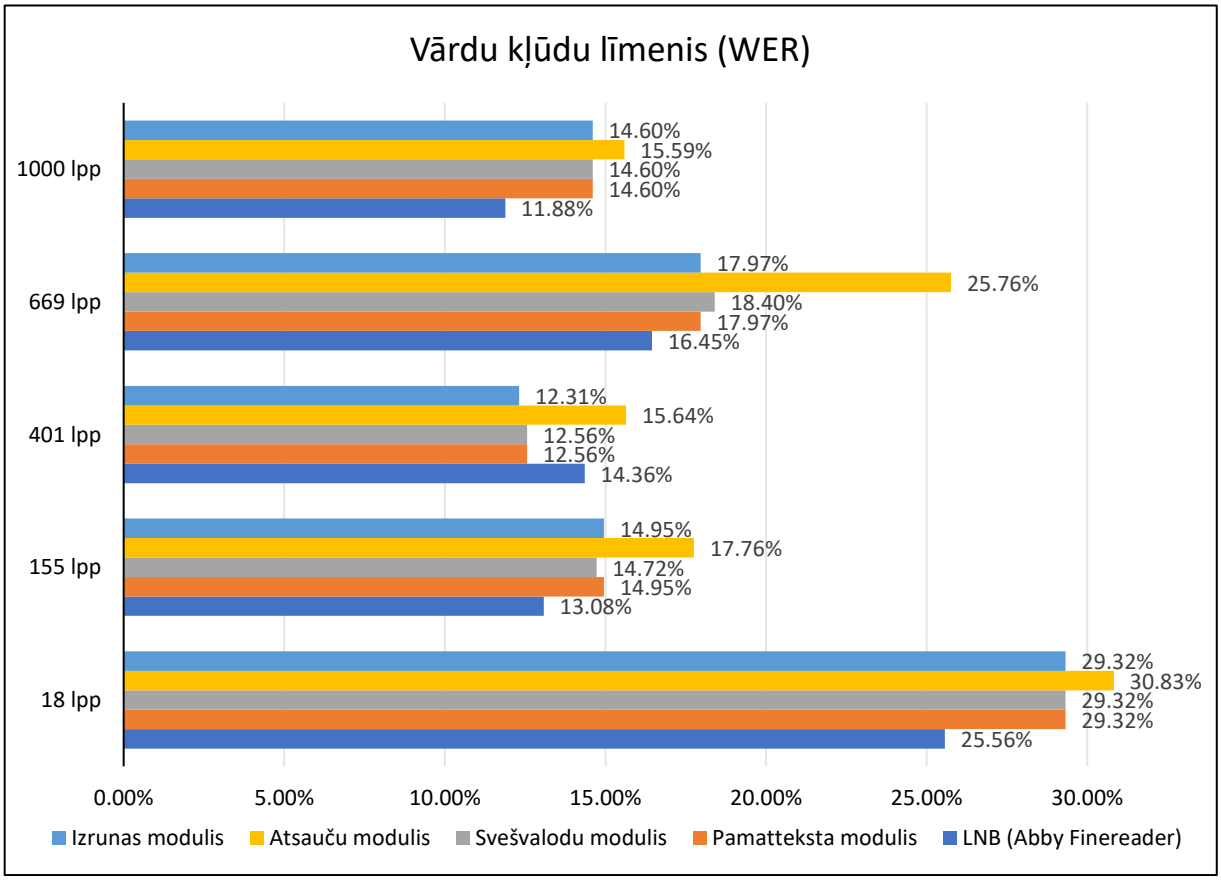
Kļūdas rezultātu iegūšanas procesā references (manuāli transkriptētais) teksts un moduļa rezultējošais teksts tika sadalīti rindiņās. Pēc tam tika veikta salīdzināšana starp katru rindiņu, izmantojot Levenšteina attāluma metodi kā pamata salīdzināšanas mehānismu (sk. att. 6.1. 6.2. 6.3.).



Att. 6.1. Vidējais CER un WER kļūdu līmeņa grafiks

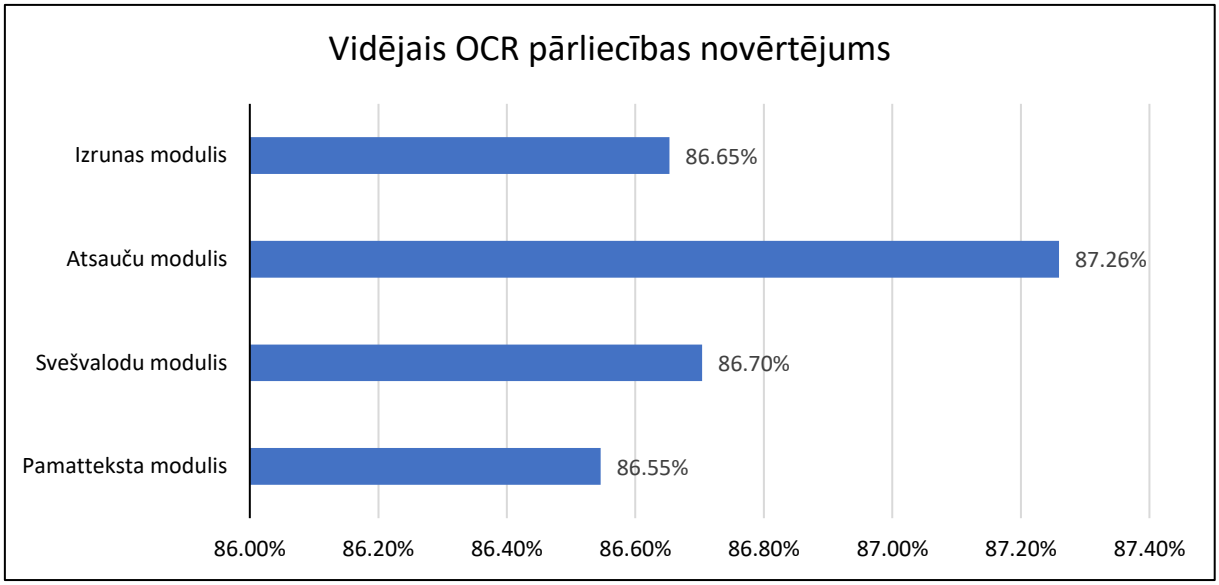


Att. 6.2. Rakstzīmju kļūdu līmeņa (CER) grafiks, katram modulim pret katru lapaspusi.



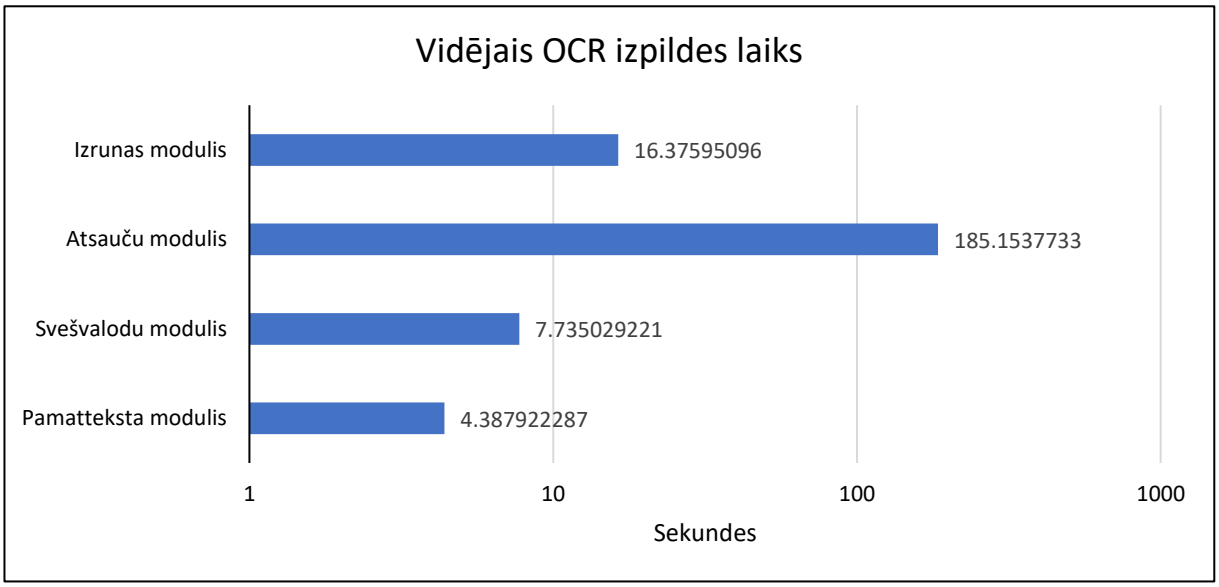
Att. 6.3. Vārdu kļūdu līmeņa (WER) grafiks, katram modulim pret katru lapaspusi.

Lai iegūtu pilnīgu priekšstatu par Tesseract rīka pārliecības novērtējuma izmaiņām un to, kā izstrādātie moduļi ietekmē šo novērtējumu, tika veikta datu apkopošana par vidējo pārliecības novērtējumu, apstrādājot katru references lapaspušu attēlu. (sk. att. 6.4.).



Att. 6.4. Vidējais OCR pārliecības novērtējums moduļu izpildē grafiks

Moduļu veiktspējas izpētes nolūkos tika veikts rūpīgs references lapaspušu attēlu apstrādes laika aprēķins katram modulim. Šajā procesā tika ņemti vērā visi apstrādes posmi, sākot no attēla iegūšanas un segmentācijas līdz vārdu atpazīšanai un rezultātu apkopošanai. Katram references attēlam tika mērīts laiks, kas vajadzīgs, lai izpildītu visus nepieciešamos uzdevumus, un šie laiki tika apkopoti, lai iegūtu vidējo apstrādes laiku (sk. att. 6.5.).



Att. 6.5. Vidējais OCR moduļu izpildes laika grafiks

Pamatteksta moduļa analīze

Pamatteksta modulis, salīdzinot ar citiem izstrādātajiem moduļiem, ir visprecīzākais rakstzīmju atpazīšanā (sk. att. 6.1.). Neatkarīgi no tā ka modulis ir precīzākais rakstzīmju atpazīšanā no izstrādātajiem moduļiem, tas nespēj precīzāk atpazīt ne vārdus, ne rakstzīmes par LNB pieejamo risinājumu, taču atsevišķos gadījumos tas spēj atpazīt labāk kā rakstzīmes (401. lpp., 669. lpp.)(sk. att. 6.2.), tā vārdus (401. lpp.)(sk. att. 6.3.).

Neskatoties uz moduļa rakstzīmju precizitāti pārējo izstrādāto moduļu salīdzinājumā, tam ir viszemākais vidējais pārliecības novērtējums no visiem izstrādātajiem moduļiem (sk. att. 6.4.), kas liecina par pārliecības novērtējuma un teksta atpazīšanas precizitātes korelācijas neesamību.

Pamatteksta modulis ir arī visātrākais izpildes laikā (sk. att. 6.5.). Tomēr ir jāuzsver, ka visi citi moduļi ir atkarīgi no pamatteksta moduļa izpildes beigām, lai iegūtu nepieciešamos datus pēcapstrādes procesam. Tas nozīmē, ka citi moduļi gaida, kamēr pamatteksta modulis pabeidz savu darbību, pirms tie var sākt savu darbu un apstrādāt datus.

Modulis spēj arī precīzi segmentēt tekstu pēc *alto* struktūrfaila (sk. att. 5.1.)

Izrunas moduļa analīze

Izrunas modulis, salīdzinājumā ar citiem izstrādātajiem moduļiem ir visprecīzākais vārdu atpazīšanā (sk. att. 6.1.), tas spēj labot pamatteksta modulī radušās kļūdas. Modulis atsevišķos gadījumos spēj vislabāk atpazīt kā vārdus, tā rakstzīmes (401. lpp.), taču tas arī ievieš jaunas kļūdas (155. lpp. 669. lpp.)(sk. att. 6.2. 6.3.).

Neatkarīgi no tā ka modulis ir precīzākais vārdu atpazīšanā no izstrādātajiem moduļiem, tas nespēj veikt nepieciešamos izrunas labojumus, kas sastopami LNB pieejamajā risinājumā. Šis ir skaidrojams ar to, ka *Tesseract* OCR rīks nav apmācīts uz šādiem specifiskiem latviešu vārdiem, kas satur fonētiskā alfabēta rakstzīmes, kā arī tas nespēj risināt šo problēmu ar tam pieejamajām svešvalodu konfigurācijām.

Modulis spēj atpazīt visus izrunas apgabalus, taču tas arī atpazīst citus apgabalus, kas nav izrunas apgabali (sk. att. 5.2.). Tas liecina, ka *Tesseract* rīks, nespēj nepārprotami atpazīt kvadrātiekavas.

Svešvalodu moduļa analīze

Svešvalodu modulis, salīdzinājumā nevienā no gadījumiem nespēj labot pamatteksta modulī pieļautās kļūdas, tas tās tikai ievieš. Tas liecina par *Tesseract* rīka nespēju pareizi atpazīt atsevišķus vārdus bez konteksta, zinot vārda rakstības valodu.

Svešvalodu modulis savā izpildes laikā ir arī visātrākais (sk. att. 6.5), jo tam ir jāatpazīst vārds skaidri definētā valodā, un nav jāveic visu valodu pārlase, kā izrunas vai atsauču modulim.

Modulis spēj precīzi atpazīt un segmentēt svešvalodu apgabalus (sk. att. 5.3.). Att. 5.3. ir novērojams (šķirklis "būda", 1. rinda), ka ir segmentēts vārds "vai" krievu valodas apstrādei, taču šis ir īpašs gadījums, kurš netiek apskatīts šajā darbā.

Atsauču moduļa analīze

Atsauču modulis, salīdzinot ar pārējiem izstrādātajiem moduļiem ir visneprecīzākais, kā rakstzīmju, tā vārdu atpazīšanā (sk. att. 6.1.). Šis modulis neuzlabo pamatteksta moduļa rezultātu, tas to tikai pasliktina. Tas ir tādēļ, ka atsauču rindkopa ir satur saīsinājumus un ciparus, ko *Tesseract* OCR rīkam ar valodas konfigurāciju ir sarežģīti atpazīt, vārdu rakstzīmju konteksta trūkuma dēļ.

Atsauču modulim ir arī augstākais vidējais OCR pārliecības novērtējums (sk. att. 6.4.), neskatoties ka tas ir visneprecīzākais, kas liecina par pārliecības novērtējuma un teksta atpazīšanas precizitātes korelācijas neesamību. Augstais pārliecības novērtējums ir skaidrojams ar faktu, ka modulis apstrādā visvairāk vārdu vienību no visiem moduļiem un tas veic pilno valodu pārlasi katram vārdam, tiecoties uz augstāko pārliecības vērtējumu.

Modulis ir arī vislēnākais savā izpildē no visiem moduļiem (sk. att. 6.3.). Tas ir tādēļ ka modulim salīdzinājumā ar pārējiem moduļiem ir jāapstrādā visvairāk vārdi, izmantojot pilno valodu pārlasi.

Atsauču modulis veiksmīgi spēj segmentēt atsauču apgabalus vārdnīcas pamatnodaļas (sk. att. 5.4.). Attēlā 5.4. pēdējais atsauču apgabals nav pilnīgi segmentēts, jo *Tesseract* ir problēmas precīzi segmentēt lapas izkārtojumu.

SECINĀJUMI

Bakalaura darba ietvaros tika veikta rūpīga izpēte par OCR tehnoloģijas darbības principiem. Autors veltīja laiku, lai iepazītos ar šo tehnoloģiju, tā darbību un iespējām. Tika apskatīti dažādi pieejamie OCR rīki, kas piedāvā atpazīt tekstu no attēliem vai skenētiem dokumentiem.

Pēc detalizētas izpētes un salīdzinājuma darba izstrādes gaitā tika izvēlēts *Tesseract* OCR rīks kā labākais risinājums bakalaura darba mērķiem. Šis rīks tika novērtēts tādēļ, ka tas ir plaši izmantots un atzīts kā viens no vadošajiem OCR risinājumiem. Tesseract OCR nodrošina spēju atpazīt tekstu no attēliem un ir atvērtā koda rīks, kas padara to pieejamu un pielāgojamu.

Bakalaura darba gaitā tika veikta rūpīga svešvalodu tekstu apgabalu analīze, kas ietvēra pamattekstu, izrunu, svešvalodu pazīmējumus un atsauces. Šie teksta apgabali tika izvēlēti un implementēti, ņemot vērā K. Karuļa *Latviešu etimoloģijas vārdnīcas* iekārtojumu un struktūru.

Darba ietvaros tika veikta *Tesseract* OCR rīka pēcapstrādes adaptera izstrāde, kas tika veikta, ņemot vērā iepriekš analizētos svešvalodu tekstu apgabalus. Šis adapteris tika veidots ar mērķi uzlabot *Tesseract* rīka rezultātus, nodrošinot labāku precizitāti svešvalodu alfabētu vārdu atpazīšanā.

Izstrādātais adapteris ļāva veikt papildus apstrādi un analīzi pēc *Tesseract* OCR rīka darbības, izmantojot identificētos svešvalodu apgabalus, kas sniedza vērtīgu kontekstu un norādes par pareizu teksta atpazīšanu. Adapteris tika pielāgots, lai ņemtu vērā vārdnīcas struktūru, rakstzīmju izkārtojumu un citus specifiskus faktorus, kas varētu ietekmēt OCR rezultātus.

Darbā izstrādātais adapteris, lai gan nespēja panākt precīzākus rezultātus salīdzinājumā ar LNB pieejamo risinājumu, sniedza nozīmīgu ieguvumu, proti, labāku izpratni par *Tesseract* OCR rīku un tā pielietojuma iespējām un ierobežojumiem.

Izstrādātā adaptera analīzē tika atklāts:

- *Tesseract* rīka pārliecības novērtējums nav patiess vērtējums atpazītā vārda precizitātei, starp tiem nepastāv korelācija;
- *tesseract_best* valodu kategorijas modeļi nav piemērotas specifiskajam svešvalodu un fonētiskā alfabēta tekstam, kas ir *Latviešu etimoloģijas vārdnīcā*;
- pilna *tesseract_best* pieejamo valodu pārlase, lai atpazītu individuālu vārdu, nav efektīva vārda atpazīšanas pieeja.

Turpmākajiem pētījumiem ir būtiska nozīme, lai sekmētu *Latviešu etimoloģijas vārdnīcas* digitalizāciju, un autors uzskata, ka ir nepieciešams izstrādāt speciālus *Tesseract* rīka valodas modeļus, kas būtu piemēroti izrunas, svešvalodas un atsauču tekstu apgabaliem.

Izstrādājot speciālus valodas modeļus *Tesseract* rīkam, kas būtu pielāgoti šiem konkrētajiem teksta apgabaliem, varētu uzlabot atpazīšanas precizitāti un nodrošināt labākus rezultātus *Latviešu etimoloģijas vārdnīcas* digitalizācijai. Izrunas teksta apgabals, svešvalodu teksta apgabals un atsauču teksta apgabals prasa specifisku pieeju un valodas apstrādes modeļu izstrādi, kas ņemtu vērā šo vārdnīcas specifiku un sarežģītību.

IZMANTOTĀ LITERATŪRA UN AVOTI

1. Roberts Brants *Optiskas rakstzīmju pazīšanas (OCR) tehnoloģiju pielietojums latviešu etimoloģijas vārdnīcām*. LUDF kursa darbs
2. Karulis, Konstantīns, *Latviešu etimoloģijas vārdnīca*. Rīga: Avots, 2001. LNB digitālā bibliotēka [tiešsaiste]. [skatīts 29.05.2023] Pieejams: http://gramatas.lndb.lv/periodika2-viewer/?lang=fr#issue:698429
3. *LNB digitālā bibliotēka* [tiešsaiste]. [skatīts 29.05.2023] Pieejams: http://gramatas.lndb.lv/#mainPage:
4. *International Phonetic Alphabet* [tiešsaiste]. [skatīts 29.05.2023] Pieejams: https://en.wikipedia.org/wiki/International_Phonetic_Alphabet
5. *Tesseract supported languages* [tiešsaiste]. [Skatīts 29.05.2023] Pieejams: https://tesseract-ocr.github.io/tessdoc/Data-Files-in-different-versions.html
6. *Alto standart* [tiešsaiste]. [Skatīts 29.05.2023] Pieejams: https://www.loc.gov/standards/alto/
7. *How Does Optical Character Recognition Work* [tiešsaiste]. [Skatīts 29.05.2023] Pieejams: https://www.baeldung.com/cs/ocr
8. Kenneth Leung *Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER)* [tiešsaiste]. [Skatīts 29.05.2023] Pieejams: https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510
9. Gidi Shperber *A gentle introduction to OCR* [tiešsaiste]. [Skatīts 29.05.2023] Pieejams: https://medium.com/towards-data-science/a-gentle-introduction-to-ocr-ee1469a201aa
10. *Python: OCR for PDF or Compare textract, pytesseract, and pyocr* [tiešsaiste]. [Skatīts 29.05.2023] Pieejams: https://medium.com/@winston.smith.spb/python-ocr-for-pdf-or-compare-textract-pytesseract-and-pyocr-acb19122f38c
11. Lucas Nogueira *Extracting Information from Images with OCR and Pytesseract* [tiešsaiste]. [Skatīts 29.05.2023] Pieejams: https://medium.com/@lucasnogsousa/extracting-information-from-images-with-ocr-and-pytesseract-ebd90c1617a4

Pielikumi

1. pielikums. *Tesseract* pēcapstrādes adapteris

Repozitorija: https://github.com/gurkitis/pytessV2.git