

Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Ökoloogia ja maateaduste instituut
Geograafia osakond

Magistritöö geoinformaatikas ja kartograafias (30 EAP)

Talunimede tuvastamise automatiseerimine Eesti põhikaartide näitel

Maria Anttila

Juhendajad:

PhD Raivo Aunap

PhD Siim Orasmaa

PhD Ardi Tampuu

Tartu 2026

Annotatsioon

Talunimede tuvastamise automatiseerimine Eesti põhikaartide näitel

Kohanimesid on traditsiooniliselt kogutud uurijate ja huviliste tarbeks kaartidelt käsitsi, kuid arvutitehnika arenedes on võimalik nimekorjet automatiseerida ja seega protsessi kiirendada. Keskseks automatiseeritud lahenduseks on masinõppele toetuv tekstituvastus (OCR – *Optical Character Recognition*). Kui nimi on olemas kohanimede loendis, saab tuvastustulemuse ühendada selle alusel. Eesti talunimede loendit aga olemas pole, kuigi praegused talunimed, mis on hajaasustusaladel katastriüksuste lähiaadressid, on võimalik ruumipäringuga kätte saada. Käesolevas töös katsetatakse tekstituvastuse mudelite sobivust eestikeelse kaardiandmestiku talunimede tuvastamisele kolme erineva aja Eesti põhikaartide näitel viiel testalal. Taolise andmestiku väljakutsed on nii keel, mis pole OCR-lahenduste peavool, kui ka kaartide müra. Tuvastuseks valiti viis vabavaralist ja kolm mitmemodaalset mudelit ning lisaks loodi vabavaralisele EasyOCR-ile ja Gemini 3-ele põhinev kombineeritud meetod. Adekvaatseid tulemusi saavutati EasyOCR-i ja kõikide mitmemodaalsete mudelite ning kombineeritud meetodiga, mis ühendab kahe mudeli tugevused. Sobivuse mõõdikuks on F1-skoor, mis on tuvastuse saagise ja täpsuse harmooniline keskmine. Lisaks uuriti võimalusi ühendada erinevate aegade nimekihte teisenduskauguse abil, kuid see nõuaks lisaks teist keeletehnoloogilist lahendust tulenevalt nimemuutuste mitmekesisusest. Töö praktiliseks väljundiks on tervet tuvastusprotsessi hõlmav Python-skript.

Märksõnad: toponüümika, talunimed, OCR, scene text recognition, mitmemodaalsed mudelid, Eesti põhikaart

CERCS kood: H370 Onomastika

P176 Tehisintellekt

P510 Füüsiline geograafia, geomorfoloogia, mullateadus, kartograafia, klimatoloogia

T111 Pilditehnika

Annotation

Automated recognition of farmstead names on Estonian Basic Maps

Place names have traditionally been collected from maps manually for research purposes, but with current computational capacity, it is possible to automate the collection of names and thus speed up the process. An essential automated solution is Optical Character Recognition (OCR) based on machine learning. If a place name is included in a gazetteer, the recognition result can be reconciled against it. However, there is no known gazetteer on Estonian farmstead names, although current farmstead names, which are the names of cadastral parcels on dispersed settlement areas, can be obtained through a spatial query. In this work, the usability of text recognition models for farmstead name recognition in the Estonian-language map dataset is tested on three editions of Estonian Basic Maps on five test sites. The challenges of such a dataset stem both from the Estonian language, which is not mainstream for OCR solutions, and from the noise on the maps. Five freeware and three multimodal models were selected for recognition, and in addition, a combined method based on EasyOCR freeware and Gemini 3 was created. Adequate results were achieved with EasyOCR, all multimodal models, and the combined method, which harnesses the strengths of the two models. In this thesis, the used suitability metric is the F1 score, i.e. the harmonic mean of precision and recall. In addition, possibilities to reconcile name layers from different times using edit distance were investigated, but this would require an additional language technology solution due to the diversity of name changes. The practical output of the work is a Python script covering the entire recognition workflow.

Key words: toponymy, farmstead names, OCR, scene text recognition, multimodal models, Estonian Basic Map

CERCS: H370 Onomastics
 P176 Artificial intelligence
 P510 Physical geography, geomorphology, pedology, cartography,
 climatology
 T111 Imaging, image processing

Sisukord

Sissejuhatus.....	5
1. Teoreetiline ülevaade.....	8
1.1 Talu defineerimise problemaatikast.....	8
1.2 Talunimedede uurimine ja allikad – keele- ja kultuuriajalooline taust.....	9
1.3 Kaardikirjade tuvastusmeetodite teoreetiline taust	14
1.3.1 Tuvastusmeetodite tööpõhimõtte ja areng.....	14
1.3.2 Kaardid tuvastusmaterjalina.....	19
2. Andmed ja metoodika	23
2.1 Praeguste talunimedede leidmine	23
2.2 Kaartide valik ja iseloomustus	25
2.3 Testalade valik ja iseloomustus.....	28
2.4 Andmete eeltöötlus, testide läbiviimine ja tulemuste järeltöötlus	30
2.4.1 Andmete eeltöötlus	30
2.4.2 Testide läbiviimine.....	35
2.4.3 Tulemuste järeltöötlus.....	37
3. Tulemused ja arutelu.....	40
3.1 Testide tulemuste hindamine	40
3.1.1 Sobivate mudelite tulemuste hindamine	41
3.1.2 Mittesobivate mudelite tulemuste hindamine	55
3.2 Nimekihistute kokkuviiamise hindamine	64
3.3 Metoodika hinnang ja kitsaskohad.....	68
3.4 Edasised uurimissuunad ja järeldused.....	74
Kokkuvõte.....	77
Summary	79
Tänuavaldused	81
Kasutatud kirjandus	82
Lisad.....	89
Lisa 1. Tehisaru kasutamise aruanne	89
Lisa 2. Tasuliste mudelite viip	89
Lisa 3. Kombineeritud meetodi viip	90
Lisa 4. Tuvastusskript	90
Lisa 5. Töövoog	91
Lisa 6a. Tulemuste F1-skoorid teisenduskaugusega 0	92
Lisa 6b. Tulemuste F1-skoorid teisenduskaugustega 0 ja 1	94

Sissejuhatus

Kuidas teada saada, mis nimesid teatud alal esineb ja ajalooliselt on esinenud? Traditsiooniliselt on nimesid kogutud käsitsi kaartidelt ja kaartide andmeid on täiendatud ja täpsustatud erinevatest ajaloolistest dokumentidest nagu hingeloendid ja vakuraamatud (vt nt Pall, 1977). Lisaks on viidud läbi välitöid, mille käigus on kasu saadud kohalikest keelejuhtidest (vt nt Saar, 2008a). Kohalike nimetundjate vajadust kinnitavad ka Kohanimenõukogule korduvalt laekuvad avaldused, milles tuuakse välja, et kaardil esinev nimi ja kohalik kõnepruuk erinevad üksteisest ja soovitakse saada nõukogu otsust ametliku (rööp)nime kohta. Süstemaatiline kohanimedele korje algas Eestis 1920ndatel, kuid siis piirduti enamasti nime häälduskuju kirjanekuga ja asukohatäpsuseks piisas külast (Kallasmaa, 1996). Taoline nimekorje on täpne, kuid väga töömahukas, ja sellepärast on piiratud üksikute piirkondadega. Arvutite arvutusvõime ning masinõppe lahenduste arenedes on aga võimalik nimekorjet automatiseerida ja seega saada kogutud suurema ala, näiteks terve Eesti riigi, kohta hajali erinevates allikates asuvad nimed ühte andmebaasi.

Teksti tuvastamine piltidelt ehk tärgtuvastus (OCR, *Optical Character Recognition*) ei ole ainult viimaste aastate teema – patenti on taotletud aastal 1929 (Wang et al., 2023; Tauschek, 1929/1935) ja ka uurimistöodes on see olnud vahendiks juba pikemalt (nt de Mello & Lins, 1999), kuigi varajaste tuvastusmeetodite täpsus oli nõrk (Wang, 2023). On tehtud võrdlusi meetodite sobivusest erinevate dokumenditüüpide jaoks (nt Eesti Kirjandusmuuseum, 2024), milles on siiski keskendutud üsna sirgjoonelisele puhta taustaga dokumenditekstile. Võrreldes sirgerealise ning puhtataustalise tekstiga on sisutihedad kaardid siiski keerulisem andmeallikas. Igal kaardil on oma eripära ja eriti vanemate kaartide puhul võivad isegi erinevad kaardilehed olla erineva väljanägemise ja kvaliteediga. Kaartide puhul ongi õigem tärgtuvastuse asemel rääkida loomulikust keskkonnast teksti tuvastamisest (*Scene Text Recognition, STR*), milles tekst esineb osana suuremast tervikust, näiteks sildid tänavapildis (vt nt Yang et al., 2022). Tuvastusprotsess koosneb olemuslikult kahest alamülesandest: tekstide leidmisest pildi pealt (*detection*) ning nende lugemisest ehk tuvastamisest (*recognition*) (nt Wang et al., 2023). Neid alamülesandeid võivad erinevad tööriistad lahendada üksteise järel, tuvastades esmalt teksti sisaldavad alad, või koos, andes mõlemad väljundid korraga. Käesolevas töös viidatakse mõistega tekstituvastus tervele, mõlemat alamülesannet sisaldavale protsessile, kui pole teisiti välja toodud.

Teksti tuvastamises pildilt on oluline lisaks sellele, mis kirjas on, ka see, kus tekst paikneb. Kaartide puhul oodatakse vastust tavaliselt koordinaatides. Selle aspekti lisandumine teeb tekstituvastuse kaartide pealt omakorda eriliseks. Mõned vanemad tuvastuse tööriistad nagu EasyOCR suudavad töödelda GeoTIFF-faile ja säilitavad koordinaatide info kõrge täpsusega, kuid uuema põlvkonna lähenemiste hulka kuuluvaid keelemudeleid (*LLM, Large Language Model* ning selle edasiarendus koos masinnägemise võimekusega ehk *VLM, Vision-Language Model*, ka VL-mudel) pole optimeeritud asukohateabe täpse tuvastamise ning väljundina andmise jaoks (vt Bordes et al., 2024; Tong et al., 2024). Mitmemodaalne mudel (*MLLM, Multimodal Large Language Model*) on keelemudel, mis suudab töödelda erinevaid sisenditüüpe (Yin et al., 2024). Käesolevas töös kasutatakse keelemudelit ja mitmemodaalset mudelit sünonüümsetena, kuigi keelemudeli tähendus on pisut teine, kuid antud töös tähtsaks on selle võimekus tuvastusel.

Sõnakasutuse seisukohast on oluline eristada ka meetodit, mudelit ja algoritmi. Kirjanduses (nt Wang et al., 2023; Long et al., 2021) kasutatakse meetodit ja mudelit sünonüümselt: mõlemad on tekstituvastuse lähenemisi, kuid Wang et al. (2023) lisavad sünonüümiks veel ka algoritmi. Antud töös kasutatakse meetodit tekstituvastuse ja selle lähenemiste tähenduses ning mudelit konkreetse tuvastustööriista kohta. Algoritm omapoolt on mudeli valem(id), millega kasutaja kokku ei puutu.

Kaartidele kantud või kantavaid kohanimesid on vaja kartograafidele, kes peavad vastama küsimusele, millised kohanimed on olulisemad, aga ka planeerijatele näiteks kultuuriväärtust omavate kohtade märkamiseks ja omavalitsustele mh kohanimede määramisel seotud otsuste tegemiseks. Nende ajalooline mõõde aga huvitab ajaloolasi ja ka kodu- või perekonnaajaloo uurijaid. Kohanimedest eriti talunimed omavad viimase puhul väärtust. Talunimede uurimine aitab leida seaduspärasusi koha- ja isikunimede vahel ja nendes võib ka säilida selliseid keele elemente, mis üldkeelest on juba kadunud (vt Harvalik, 2004).

Kohanimede kohta leidub erinevaid loendeid ja kogusid nagu Kohanimeregister (AKS-KNR) ja Kohanimeandmebaas (KNAB) ning rahvusvaheliselt näiteks GeoNames. Erinevates andmekogudes on nimesid erinevatelt haldusjaotuse tasanditelt, kuid Eesti talunimesid on autori teada kogutud digitaalsel kujul ainult Kohanimeandmebaasi ning Evar Saare doktoritöö (2008a) raames valminud Ajaloolise Võrumaa kohanimede andmebaasi (AVKA), milles talunimed sisalduvad osana asustusnimedest. Talunimesid ongi kogutud osana suuremast korjest, mitte eraldi neile keskendudes.

Käesolev magistritöö aga keskendub just Eesti talunimedele. Töös katsetatakse erinevaid masinõppele toetuvaid arvutite sõnetuvastuse võimalusi Eesti põhikaartide näitel 1990ndatest aastani 2026. Erinevaid lähenemisi on üheksa (vabavaralised mudelid EasyOCR, MMOCR, NewOCR, PaddleOCR ja Tesseract, tasulised mudelid ehk kolm suurt mitmemodaalset mudelit Gemini 2.5 Pro, Gemini 3 Flash ja Qwen3-VL-32B ning kuna ükski ei osutunud tervele andmestikule sobivaks, loodi kombineeritud meetod, milles tekstid leitakse EasyOCR-i ja loetakse Gemini 3-ega). Osa mudelite puhul aga kaardistatakse ainult nende võimekust kaartidelt nimede tuvastamisel ilma tulemusi kaardile kandmata. **Magistritöö eesmärgiks on võrrelda erinevate mudelite võimekust eestikeelse, kohati murdelisi jooni sisaldava kaardiandmestiku peal ja sobiva(te) mudeli(te) alusel välja töötada meetodika talunimede tuvastamiseks rasterkaartidelt. Töö praktiliseks väljundiks on tervet tuvastusprotsessi hõlmav Python-skript.** Uurimuse aluseks on viis testala suuruses 5x5 km ehk praeguse 1:10 000 Eesti põhikaardi lehele vastav ala koos 200 m puhvriga. Tekstis viidatakse mõistega kaardileht ühele Eesti põhikaardi lehele sõltumata mõõtkavast ja mõistega mosaiikkaardileht lõigatud ja kokkukleebitud rastrite kogumile. Testalade koguse valikul on lähtutud sellest, et kaardilehed on erineva taustainfo, kuid ühtlase kvaliteediga.

Lisaks tuvastusele lisandub nimede ajalise kokkuviimise problemaatika, sest mõistmaks nimede arengut on vaja omavahel siduda sama objekti nimi erinevatel aegadel. Pikemas ajaloolises perspektiivis saab täheldada eesti keele ortograafia muutuseid, kuid uuemal ajal seisnevad paljud muutused hargtäiendite kasutamises – näiteks vanem talu on saanud oma nime ette hargtäiendi *Vana-*, kui lähestikku on kerkinud uuem taluhoone.

Eesti põhikaardid kolmes erinevas seerias (aasta 2026 põhikaart, 10-tuhandeline digitaalne kaart ning 20-tuhandeline trükikaart) on käesoleva töö näidismaterjal, kuid sama töövoogu saab üldjoontes laiendada ka vanematele ajaloolistele kaartidele. Ajalooliste kaartide puhul

lisanduvad mõningad väljakutsed, mida käsitletakse töös teoreetilisel tasandil. Töö on tugevalt geoinformaatiline ja seega fookusest jääb välja nimede etümologiseerimine. Tegemist on interdistsiplinaarse ehk valdkondadevahelise tööga, mis ühendab (geo)informaatikat, keeleteadust ja keeletehnoloogiat ning ajalugu.

Erinevaid mudeleid ja algoritme, mille abil digitaliseeritud dokumentide pealt nimesid kätte saada, on väga palju ning nende areng on kiire. Käesolevas töös ei püüta otsida algoritmide hulgast kõige parimat, vaid eesmärgiks on leida piisavalt hea tulemuslikkusega mudel, mille kasutamine oleks lihtne ka programmeerimisoskusteta inimestele. Eesmärgiks on hoida töövoog võimalikult lihtne, et selle skriptide käivitamine ei nõuaks sügavat informaatika arusaama, suurt arvutusvõimsust ega keerulist installeerimist. Töös katsetatakse ja hinnatakse autori teada esimest korda erinevaid tekstituvastusmudeleid Eesti kaardiandmestiku peal.

Lisaks otsitakse töös vastuseid järgmistele küsimustele:

Mis on testitud mudelite võimekuse piirid eestikeelse kaardimaterjali automaattuvastusel?

Kas automaatne tekstituvastus sobib ka eestikeelsele ning kihistunud ja murdelisi elemente sisaldavale talunimestikule?

Kuivõrd teisenduskaugus sobib ajaliste nimekihistute kokkuviimiseks?

Kõik nimed, nii need, mis töös kasutatud kaartidel esinevad, kui ka need, mis on mudeli poolt tuvastatud, on töös esitatud kaldkirjas, näiteks *Ando*. Jutumärkides on esitatud teistest allikatest pärit näidiseid, mis ei esine analüüsitud kaartidel ega pole mudelite tuvastatud, näiteks „Sooääre“. Tuvastatud nimed on esitatud mudeli antud kujul ja sellepärast esineb nendes komponente, mis ei vasta eesti keele ortograafia reeglitele (näiteks väike algustäht, erinevad kirjavahemärgid sõne sees jpt). Lisas 1 on toodud tehisarude kasutamise aruanne.

1. Teoreetiline ülevaade

1.1 Talu defineerimise problematikast

Selle üle, mis üldse on talu ja kuidas seda piiritleda, saab diskuteerida, ja definitsioon sõltubki uurimuse otstarbest ja fookusest. Tihtipeale võetakse seda enesestmõistetavalt ja uurimustes tuuakse eraldi definitsioon välja pigem harva, kuid näiteks Gea Troska (1995) täpsustab, et talu on üheperemajapidamine.

Käesolevas töös defineeritakse talu lähtuvalt Maa- ja Ruumiameti andmestikust: Kohanimeregistri kirjapanekust ja seosest hoonega Eesti topograafia andmekogus (ETAK). Nimede kättesaamist on põhjalikumalt kirjeldatud peatükis 2.1 Praeguste talunimedele leidmine, kuid lühidalt olgu öeldud, et antud töö kontekstis talu on maaüksus, mille tunnuspunkt asub hoone sees. See vastab katastriüksuse lähiaadressile, mis on ka Eesti põhikaardi 1:10 00 leppemärkide seletuses (Maa- ja Ruumiamet, 2025a) välja toodud. Praeguseid talunimesid on Eestis 106 795 tk. Vanemate kaartide puhul lähtutakse talu defineerimisel kaardil kasutatud kirjastiilist. (Talunimedele eristuvust ülejäänud nimestikust selgitatakse peatükis 2.2 Kaartide valik ja iseloomustus.)

Ei ole üheselt arusaadav, millal tegemist on taluga, ja kaasates erinevaid allikaid võib jõuda erinevate tulemusteni. Eesti keele seletava sõnaraamatu [EKSS] (2009) järgi talu on „põllumajanduslik väikemajand, mis hõlmab maa koos sellel olevate hoonetega” ning näitena liitsõnast on toodud muuhulgas turismitalu. See tekitab küsimuse, kas turismitalu tegevusvaldkonda peab tingimata kuuluma põllumajandus, et ta saaks olla põllumajanduslik väikemajand, ja kas siis talunimed tuleks otsida kaasates maatulundusmaa andmestik. Näiteks Aleksander Pertelson (2025) on oma bakalaureusetöös tekitanud talunimedele andmestiku valides katastriüksused, mille sihtotstarve on maatulundusmaa ning mille sees asub elu- või ühiskondlik hoone.

Samasugust tegevusvaldkonnaga seotud kontroversiaalsust esindab päring Äriregistrist (2025, seis 9/25), mis annab 4002 vastet, milles nimi sisaldab sõne „tal” ning mis on staatusega „Registrisse kantud”. Osal nendest on aga aadress selgelt linnalisel alal näiteks Tallinna või Tartu kesklinnas. Siiski on ilmne, et aadressikoht ja tegelik tegevuskoht ei pea kokku minema ja eriti, et tänaste taluettevõtete asukoht ei pea ühtima ajalooliste talukohtadega. Lisaks Äriregistrist leidub näiteks koondettevõtteid, mille tütarettevõtetest ainult ühe tegevusala on seotud talupidamisega, või hulgimüügiga tegelevaid ettevõtteid, mis ostavad talusaaduseid kokku. Ka Maa-amet¹ (2022) on viinud läbi sama analüüsi aasta 2015 andmetel ja jõudnud järeldusele, et umbes 1/5 Äriregistri tulemustest annab sama nimega vaste Aadressiandmete süsteemist maaüksuse nimena.

Lõhet Äriregistri ning Maa- ja Ruumiameti andmestiku vahel seletab vähesel määral sisseseatud tingimus, et objekt on registrisse kantud, mis Maa- ja Ruumiameti andmestikus ei kajastu. Ilma selletagi on siiski Äriregistri tulemusi ainult 9818 (seisuga 9/25). Seletuseks saab pakkuda ka seda, et Maa- ja Ruumiameti hoonestusest vaid osa on tegutsevad ettevõtted, mis niisiis oleks Äriregistrisse kantud ja ülejäänud näiteks talu-liidesega pere kodud. Kuna käesoleva töö uurimisobjektideks on Eesti põhikaartide nimed, mis ei oma seost nüüdisaegse

¹ Aastal 2022 kasutusel olnud nimekuju.

Äriregistriga, lähtutakse töös Maa- ja Ruumiameti andmestiku väljavõttest, kuid tuleb olla teadlik suurtest erinevustest andmestike vahel ja valida andmed teadlikult vastavalt oma uurimissuunale ja -küsimustele.

Ka siis, kui on ühtne kokkuleppe selle kohta, mis on talu, tuleb nentida, et see on üldmõiste, mille alla mahuvad erinevat tüüpi talud. Näiteks mõisa ajal jagati talud ostutaludeks ja renditaludeks, maareformi käigus mõisamaad jagati tükkideks ja nendele kruntidele tekkinud majapidamised on asundustalud, nagu selgitab Ivar Kaigu (Rahvusarhiiv, 2023). Kui talul on olnud alguses teine nimi ja maareformi käigus on see saanud tunnuseks numbri, saab järeldada, et tegu on vana renditaluga (Rahvusarhiiv, 2021). Lisaks esineb keelelisi erinevusi – Pajusalu jt. (2018) toovad välja, et põhilised nimetused on lisaks mainitule talo ja pere (viimane Lääne- ja Loode-Eestis), kuid kasutusel on muuhulgas ka koht (Saaremaal), tare või tarõ (Võrumaal, eriti jagunenud talude nimedes kuid ka sünonüümselt) ja saun (popsi- ja saunikukohtades) ning järelkomponendina on leida -aseme, -elu (Saaremaal), -rahva (Virumaal ja Põhja-Tartumaal) ning -toa ja -õue, millest viimane viitab sellele, et talu on vahepeal olnud tühi või kadunud.

1.2 Talunimede uurimine ja allikad – keele- ja kultuuriajalooline taust

Talunimed on omamoodi nähtus, kuna nad asuvad kohanimede (toponüüm) ja inimeste nimede (antroponüüm) ülekattealal, sest talu nimega viidatakse nii hoonetele või hoonete kogumile kui ka talu elanikele (Harvalik, 2004). Evar Saar (2008a) selgitab, et talunimede traditsioonis on kaks võimalust: 1) traditsioonilises süsteemis on kehtinud kindlad, püsivad talunimed ja elanikud on saanud endale lisanime talu järgi ja 2) talukohtasid on nimetatud nende elanike järgi, sest talunimede muutumatust pole oluliseks peetud. Sellel puhul esineb piirkondlikke erinevusi – näiteks toob Saar (2002) oma vanemas töös välja, et põlisele räpinlasele oleks keeruline selgeks teha, et talunimi peaks käima koos talukohaga, sest kohaliku arusaama järgi tähendab talunimi eeskätt peret ja alles teisejärguliselt elukohta. Sama kehtib Setomaal, kus tähtis on talunime seos elanikuga, mitte niivõrd talukohaga (Faster, 2013). Marja Kallasmaa (2001) hinnangul on talu nimi algul muutunud vastavalt omanikule või perekonnapeale ja alles hiljem on nimi kinnistunud talukohale.

Ta toob ka välja (Kallasmaa, 2023), et talunimede uurimisel tekib oluline küsimus, kas lähtutakse nimedest (nimekesksus) või kohast (kohakesksus), ja mainib, et traditsiooniline uurimissuund on keskendunud nimedele (nt Kallasmaa, 1996; Pall, 1969) ja alles hiljem on liigutud üle kohakesksusele. Ta (Kallasmaa, 2023, vt ka Laansalu, 2014; Troska, 1995) aga esindab seisukohta, et kohanimede ajaloo keelelisel uurimisel tuleks lähtuda nimekesksusest, kuna kohanimed on kaartidest vanemad ja sellepärast tihti puudub võimalus nime alusel täpne ajalooline koht ära määrata.

Talunimede kujunemise uurimise põhilisteks allikateks on lisaks vanadele kaartidele ja kohalikele keelejuhtidele ajaloolised dokumendid (Kepsu, 1987), Eestis eeskätt revisjonikirjad (hingeloendid) ja vakuraamatud (vt nt Laansalu, 2014; Pall, 1977). Hingeloenditele ja vakuraamatutele lisanduvad vähesemal määral ka teised arhiivimaterjalid näiteks Põllutöoministeeriumi Katastri ja Maakorralduse osakonna (sic!) maaüksuste nimekirjad ja boniteerimistoimikud, põllumajandusloendid, talundilehed ning nõukogudeaegsed hooneregistratoimikud ja majapidamisraamatud (Rahvusarhiiv, 2021). Andmed on aga mainitud dokumentides osaliselt olemas üksikute talude kohta ja osaliselt teatud piirkondade,

tihtipeale kihelkondade kaupa, ning kuigi dokumente on tublisti digitaliseeritud², nende läbitöötamine on eelnevast tulenevalt mahukas käsitöö ja seega tihtipeale vaadeldakse nimesid peamiselt piirkondade kaupa (vt järgmine lõik). Lisaks tuleb ajaloolises uurimises tähele panna, et vakuraamatutes ja hingeloendites on peremeeste nimed kirja pandud seoses (põllu)maade, mitte taludega, ja seoses pere kolimisega asuvad paljud juba 17. sajandist saadik sama nime kandnud talud kaks sajandit hiljem hoopis teise koha peal (Troska, 1995). Nimede sügavuti uurimist raskendab ka see, et ametlik nimi ja kogukonna kasutatav nimi ei tarvitse ühtida – näiteks Helju Linder (2007) toob Kaiust näitena, et ametliku nimega „Sooääre“ väiketalu tunneb kogukond nimega „Juhansaun“.

Talunimedele keskenduvate ülevaadete hulk kirjanduses on üsna tagasihoidlik – peamine allikas on Gea Troska (1995) ajalooline vaade, kuigi teemal on teinud ettekandeid ka näiteks Peeter Päll (2008) ja Evar Saar (2008b). Üldiselt talunimed Eestis on metoodikale vaatamata pälvinud tähelepanu peamiselt üksikutel aladel, millel on tublisti kaardistatud kohanimesid ja sealhulgas talunimesid ja nende ajaloolisi muutusi (keeajalooliselt nt Audrus Kallasmaa (2023), Kosel Laansalu (2014), Setomaal Kiristaja (2013), Nõos Ernits (2011), Kiilis Tärk (2010), Saaremaal Kallasmaa (1996) ja Põhja-Tartumaal Pall (1977³; 1969) ning kodu- ja kultuuriajalooliselt nt Kaius Linder (2007) ja Kiidjärvel Kalle et al. (2006), kuigi järgmiste fookus on nimede muutumise asemel pigem kodukoha ajaloos). Geograafias on tehtud kaks bakalaureusetööd, milles uuritakse talunimesid: Aleksander Pertelson (2025) keskendub Viljandimaa talunimedele ja nende tähendusväljadele ning Gettel Sink (2013) vaatab Valjala kihelkonna talunimesid osana nimekooslust. Troska (1995) aga nendib, et hoolimata laiaalalisest ja interdistsiplinaarsest huvist näiteks külanimede vastu, on talunimede uurimine olnud ajalooliselt üsna tagasihoidlik. Marja Kallasmaa (2023) hinnangul ongi talunimesid „raskem uurida [kui külanimesid] nende vähema ajalise ulatuse ja nimekorduste tõttu“. Lisaks eeltoodud talunimesid kaardistanud allikad on peamiselt humanitaarvaldkonna tööd, mispärast on nendes põhjalikult välja toodud nimede etümoloogiat ja kujunemislugu, ja nendest ainult Pall (1977) kaardistab erinevate lõppudega nimede ja sealhulgas ka talunimede levialasid.

Tulenevalt keeleteaduslikust lähenemisest on mustreid leitud pigem sõnade struktuurist. Põhiline Eesti kohanimede liigitus on kaheks: 1) nimed, mis nõuavad determinanti ja 2) nimed, mida saab kasutada ka determinandita (Kallasmaa, 2005). Determinant tähendab põhisõna ja see kirjeldab koha liiki: talu, küla, mägi, mõis jne. (Tärk, 2010). Esimene rühm koosneb asustusnimedest, teise moodustavad ülejäänud kohanimed (Tärk, 2010). Teine võimalus on rühmitada eestikeelsed kohanimed lingvistiliselt struktuuri järgi jagades need lihtnimedeks ja liitnimedeks (Kallasmaa, 2005). Liitnimed koosnevad vähemalt kahest tüvest, lihtnimed ühest tüvest või tüvest ja tuletusliitest (Tärk, 2010; Pall, 1977). Liitnimi on näiteks „Kanametsa“, kuid Pall (1977) mainib, et kõikide liitnimede puhul põhisõna ei kirjelda objekti liiki – ka antud näide on talunimi. Palli (1977) hinnangul taolised nimed on sekundaarsed ja saanud oma nime lähestikus paiknevalt loodusobjektilt. Nähtus on nimesiire ehk „pärisnime siirdumine mingile teisele nimeobjektile, ühelt kohalt teisele“ (OOS, 2024). Talunimedest liitnimede hulka kuuluvad ka kahest isikunimest koosnevad nimed (Pall, 1977), käesolevas töös näiteks *Larase-Andruse*. Lihtkohanimedest arvukaim rühm on elliptilised nimed, mille põhiosa on kadunud,

² Rahvusarhiivi teenuses Saaga saab tutvuda revisjonide materjalidega, <https://www.ra.ee/dgs/explorer.php>

³ Marja Kallasmaa (1995) hinnangul võtab Pall oma hilisemas uurimuses kokku terve senise Eesti kohanimeuurimise (toponomastika) ning tema järeldused on rakendatav Põhja-Tartumaast oluliselt laiemale alale.

ja elliptiliste nimede rühmas arvukaim talunimede esindus on isiku hüüd- ja lisanimedest kuid ka taime- ja loomanimetustest välja kujunenud nimedel (Pall, 1977).

Kallasmaa (2005) mainib lisaks, et Eestis on hulk hilise tekkega külasid või asunduskülasid nimedega nagu „Mäe“, „Metsa“ või „Nurme“, ja nendest rääkides tuleb alati lisada determinant, sest muidu peetakse neid talunimedeks. Samas determinandiküsimus saab sõltuda ka kontekstist: Ta toob välja, et näiteks nimede kogumisel on keelejuht arvestanud, kas küsitaja tunneb konteksti või mitte ehk väiksemate looduskohtade puhul on peetud oluliseks põhisõna nimetada kuid küla- ja talunimede puhul mitte (Kallasmaa, 2005).

Vanimate asustusnimedena tuntakse eelkõige külanimesid, kuigi talupoja põhiline majandusüksus on traditsionaalselt olnud eeskätt talu (Troska, 1995). Troska (1995) viitab Taani hindamisraamatule (*Liber Census Daniae*) aastalt 1241, milles on toodud välja Harju- ja Virumaa asulanimesid (478 tk) ilma selgituseta, mis liiki asulaga on tegu. Sellest allikast on siiski mõningaid vanu talunimesid võimalik saada kätte loogikaga: kui 13. sajandi talu normaalsuuruseks peetakse ühte adramaad (Troska, 1995; vt ka Johansen, 1933), võib Paul Johanseni (1933) arvates pidada üks kuni kolm adramaad hõlmavat üksust pigem üksikpere majapidamiseks (hajataluks) kui külaks. Troska (1995) selgitab, et alles hilisematel sajanditel sai kombeks eristada loendamisel talusid küladest ja seda eeskätt Põhja-Eestis. Põhjuseks oli soov panna paika mõisavalduste ulatus, mida tehti asustatud üksuste loendamise teel. Samas ta viitab ka 1586. aasta revisjonile, milles on iga mõisa valduses kirja pandud muuhulgas külad ja eraldiseisvad talud/pered, kuid toob välja, et külade ja hajatalude nimesid ei ole võimalik mitte mingisuguse tunnuse alusel üksteisest eristada. Lõuna-Eestis Taani hindamisraamatule sarnane allikas puudub ja talude olemasolu saab järeldada mõisavaldusi puudutavate ürikute nagu läänistuskirjade ja ostu-müügi-lepingute järgi, milles on toodud mõisale kuuluvad asulad ja nende suuruse järgi saab järeldada ka talude olemasolu (Troska, 1995).

Talunimesid peetakse külanimedest nooremateks, kuid samas on ajaloolased kindlaks teinud, et talu on vanem asustusüksus kui küla ja vanimad talunimed ulatuvad kirjanekutest kaugemale minevikku (Kallasmaa, 2023). Sellele küsimusele vastamiseks, millal talunimed on tekkinud, andmed aga puuduvad (Kallasmaa, 2001). Näiteid, kus esmamainitud talunimest või talupoja lisanimest on saanud küla või mõis, toob Audrust Marja Kallasmaa (2023), ning Valjala kihelkonnast leidub Gettel Sinki (2013) tulemustel külanimeks teisenenud talunimesid. Ka Gea Troska (1995) sõnul leidub eriti Lõuna-Eestist palju – kuid piirkondlikult – näiteid selle kohta, et kindlaks kohanimeks mõistetavaid talunimesid on esinenud juba 16. ja 17. sajandil, näiteks „Taliwer“ (Tälvvere küla) ja „Tapik“ (Tapiku mõis), ning ka Marja Kallasmaa (2001) toob välja, et suur osa 16.–17. sajandi lisanimedest on tänapäeval talunimed ning ei saa välistada, et nad poleks sama ülesannet juba toona täitnud. Talunimedeks loetakse need kas vastavalt dokumentidele, mis annavad teavet külamaade jaotuse ja taluvalduste paigutuse kohta, või osaliselt ka 17. sajandi esimesel poolel Rootsi poolt Lõuna-Eestis läbiviidud revisjoni tulemustele, milles on lisaks küladele märgitud ka üksikperesid (Troska, 1995).

Niivõrd kui talunimesid on uuritud, on huvi keskmes olnud peale üldiste mustrite pigem etümoloogiline lähenemine ehk üksikute sõnade kujunemine. Võib näiteks küsida, mis tähendusega on Lõuna-Eesti nimed „Kööri“, „Kiini“ või „Erdu“. Marja Kallasmaa (2000) toob Saaremaa näitel välja, et kombeks on olnud otsida raskesti seletatavate talunimede etümoloogiaid saksa eesnimedest ja nendest lähtunud perekonnanimedest lisades hiljem eelduse, et need käibelt kadunud isikunimed on olnud eestlaste hulgas kasutusel, kuigi allikate puudumisel pole neid dokumenteeritud (Kallasmaa, 2001). Valdek Pall (1969) omapoolt

mainib, et siiski pole väärtuseta tõdemus, et kohanime etümoloogia võib olla raskesti tuvastatav, sest see võib viidata asjaolule, et nimi on väga vana ja moondunud, selle taustaks olev üldnimi (apellatiiv) on keelest juba kadunud või taustaks olev isikunimi on meile tundmata või võõrast päritolu. Väljakutset pakkuv on ka see, et laenutatud nimed ja nende edasimugandused eesti keelde võivad olla väga mitmekesised, kuigi abiks saavad olla allikad nimede vanemate üleskirjutustega (Kallasmaa, 2001).

Üldiselt kehtib arusaam, et talupoegade lisanimedest on aja jooksul ja koos põlvkonnavahtetustega kujunenud välja esiti perekonnanimedele vastavad nimed (perenimed) ja siis talunimed (Troska, 1995). Tegemist oli pika protsessiga ja perenimede muutumine püsivateks talunimedeks sõltus suuresti nii üldistest kui ka kohalikest tingimustest, ja tulenevalt geograafiast oli areng Põhja- ja Lõuna-Eestis teistsugune: Lõuna-Eesti küngaste vahel omaette paiknenud hajatalude jaoks kinnistus perenimi talunimeks varem kui Põhja-Eestis ning sellepärast esineb Lõuna-Eestis rohkem ka vanadest eesnimedest väljakujunenud talunimesid (Troska, 1995). Vanimad talunimed on lähtunud peremeeste nimedest, kas ees- või lisanimest või mõlemast (Troska, 1995). Esimesed talunimede kirjapanekud pärinevad 16. sajandi vakuraamatutest ja revisjonikirjadest (Laansalu, 2014), kuid vakuraamatuid, milles talu on nimetatud peremehe järgi, leidub juba 13. sajandilt (Troska, 1995). Marja Kallasmaa (2023) toob siiski välja, et talunimesid hakati panema kirja peamiselt 16. sajandil, sest varem käis arvestus adramaade järgi, kuid sellest, et maksustamisalus oli teistsugune, ei saa järeldada, et talunimesid poleks kasutatud. Perenimed aga kinnistusid taludele 18. sajandi lõpuks ja 19. sajandil neid kanti juba ametlike talunimedena kaartidele ja katastriraamatutesse (Troska, 1995). Joonisel 1 on toodud Eesti Kohanimeraamatu [KNR] (2017) kirjena hilisem näide nimeahelast, nimest, mille areng on alanud isikunimest ning liikunud sellest talunimeks ja edasi külanimeks.

Nutru [ˈnutru] <-le> **Kõp** – paik (küla) Viljandi maakonnas Viljandi vallas (Väike-Kõpu mõis), 1695 *Nuttru* (küla), 1839 *Nutro*.  **A2**
• Külana 1970. a nimekirjas, liideti 1977 Väike-Kõpuga. Nime lähtekohaks on tõenäoliselt talunimi, mille aluseks isikunimi, vrd vanapõhja *Knútr*. Nimi on kasutusel olnud ka Eestis kujul *Knuter* ~ *Nuter* ~ *Nuuter*. Vrd *Nuutre*. – MK
EAA.567.3.86:4, L 3p; EAN; Insley 1994: 270; Rücker

Joonis 1. Näide nime arengust (nimeahelast).

Koha- ja täpsemalt talunimedes esineb Eestis piirkondlikke erinevusi. Marja Kallasmaa (2005) selgitab, et Eestis leidub kahte tüüpi kohanimekooslusi ehk teatud ala kohanimed, mida kohalikud elanikud kasutavad ja mis moodustavad alal tervikliku nimesüsteemi. Rannikualade ja saarte kooslus koosneb iseseisvatest tüvisõnadest moodustatud nimedest ja koosluses on palju loodus- ja viljelusnimed, mida pole teisest nimest tuletatud. Ülejäänud Eesti arvatavasti nooremat kooslust iseloomustavad küla- ja talunimed, millest on tuletatud loodus- ja viljelusnimed (Kallasmaa, 2005). Saaremaa kohanimed on tugevalt mõjutanud alamsaksa ja rootsi keeled, kuid on toodud välja ka friisi keele mõju eriti talupoegade lisa- ja ristnimedes, ja nendest nimedest ongi saanud esiti talu- ja hiljem üldisemad kohanimed (Kallasmaa, 1996). Kokkulangevust esineb eriti alamsaksa ja friisi mehenimede ning Saaremaa talunimede vahel, kuid pole siiski alati võimalik eristada, kas nimi on friisi või alamsaksa päritoluga (Kallasmaa, 2001). Võõrkeele mõju on ka Põhja-Tartumaa vene isikunimedest tuletatud talunimedes (Pall, 1977) ning kohati läti keelest mõjutatud nimedes lõunapiiri ääres (Pajusalu jt., 2018). Setomaa eripärana mainib Mariko Faster (2013), et kuigi tavapärane on olnud nimetada talu peremehe järgi, seal leidub mitmeid talunimesid, mis tulenevad naise nimest, sest perenaine on asunud

talude pidama peremehe kadumise korral. Need on osa vanast nimesüsteemist, mis kõrvaldati 1920ndatel maade kruntimise käigus ja mille tagajärjel taludele pandi ametlikud, meelevaldsed nimed. Tänapäevased ametlikud nimed ei erine suures osas Eestis üldiselt kasutatavates talunimedest, kuigi nendes on näha setokeelsust (Faster, 2013).

Valdek Pall (1977) on uurinud Põhja-Tartumaa nimesid ajaloolistes dokumentides ja toob nende alusel välja, et talunimedel on tihti seotud isiku- või isiku lisanimega ja esimesse rühma kuulub tema hinnangul suurim osa talunimedest. Siis talunimeks on saanud peremehe eesnimi. Lisanimeks on olnud elukutset tähistav nimi („Sepa“), päritolu („Narva“ Põhja-Tartumaal) või isikut iseloomustav omadus („Puhta“). Nime motiiviks võib olla ka looma või linna nimetus („Ilvese“), kuid piir lisanimega on sellel puhul hägus, sest osa elusolendite nimetustest on kasutatud ka lisanimena. Ka hüüdnimi on saanud olla nimetamise aluseks, näiteks „Noorhärä“. Siiski leidub ka teisi nimetamise aluseid nagu talu paiknemiskoht („Mäeotsa“), taluga kaasnevad hooned („Aida“) või talu ümbruses asuv loodusobjekt („Aru“), kuid viimaste puhul on piir jälle hägus, sest talunimi on võinud lähtuda ka perekonnanimest ja seos loodusobjektiga võib puududa (Pall, 1977).

Talunimedel on tihedust kaardil võib mõjutada ka hoonestuse muster. Eestis eristatakse iseloomuliku hoonestuse muustriga külatüüpe (vt nt Troska, 1998/2008; Troska & Šlõgina, 1976). Muster ei tarvitse täielikult ühtida talude paiknemisega, sest külates leidub ka hooned, mis ei kvalifitseeru taludeks, kuid üldised eripärad joonistuvad muustriest välja. Mustreid põhjustavad looduslikud tingimused nagu vööd ja Lõuna-Eesti mäed ning piirkondlikud kultuursed erinevused nagu tänavakülad venekeelsete kogukondade aladel (vt Troska, 1998/2008; Troska & Šlõgina, 1976). Looduslikud tingimused kehastuvad kaardil näiteks samakõrgusjoonte ja nende tiheduse kujul, mis omapoolt mõjutavad seda, kus kaardil leidub ruumi nime paigutamiseks ja seekaudu seost nimesildi ja nimeobjekti vahel.

Pärast Eesti taasiseseisvumist andis Kohanimenõukogu [KNN] (1995) kohalikele omavalitsustele soovitusel talunimedel ennistamise kohta. Soovituste andmise ajaks oli valminud kohanimeseaduse eelnõu, kuid kohanimekomisjoni hinnangul oli talunimedel valiku ja kinnistamise küsimus niivõrd pakiline, et selle jaoks tuli koostada eraldi juhend. Juhendi kohaselt tuleb mitme võimaliku nime puhul eelistada „kohapeal tuntumaid, samuti ajaloo ja kultuuri seisukohast väärtuslikumaid nimesid“ ning järgida kohaliku kõneviisi kirjutades nimi nii nagu teda hääldatakse (näiteks pika täishäälikuga hääldatav nimi „Ooblu“, mitte „Obolu“). Elanike perekonnanime tohib talule panna ainult erandjuhtudel, kui nimi on eestipärane, vaid pigem tuleks võtta kirjapanekutest vanem käibel olnud nimi, nimeta talu puhul tuleks eelistada mõnda lähedast loodusobjekti või uut „kohanimeilmelist“ nime. Lisaks esitatakse nõuandeid maakatastriüksuste liitmise ja lahutamise kohta. Liitmise puhul tuleb nimi eelistatavalt võtta sellelt üksuselt, kus põhielamu asub. Lahutamise puhul tuleb vana nimi jätta vanale talule ehk nn tuumosale. Muude üksuste nimed võib panna lähedaste loodusobjektide järgi, kasutades täiendit (nt „Uue“, „Vastse“) või teisendada muul kombel näiteks kasutades nimeahelat („Kerba“ talu lisatalu metsa läheduses võib kanda nime „Kerbametsa“). Ka täiesti uus kohanimeilmeline nimi on lahutamise puhul lubatud (KNN, 1995). Talude nummerdamist („Põllu 1“, „Põllu 2“) tuleb aga vältida (KNN, 1995) ja praegu selle keelabki kehtiv Kohanimeseadus [KNS] (2003 § 13, lg 5).

1.3 Kaardikirjade tuvastusmeetodite teoreetiline taust

Teksti tuvastamiseks skaneerimise või pildistamise teel omandatud pildi pealt on palju erinevaid mudeleid ja lähenemisi. Selles peatükis selgitatakse tuvastusmeetodeid üldisel tasemel. Erinevate rakenduste ja mudelite taustalgoritmide tehniline võrdlus jääb selle töö fookusest välja, kuna eesmärgiks on lihtsalt leida spetsiifilise andmestiku jaoks piisav tuvastuse töövoog.

1.3.1 Tuvastusmeetodite tööpõhimõtte ja areng

Klassikalise meetodina automatiseeritult teksti tuvastamiseks peetakse (vt nt Wang et al., 2023; Chiang & Knoblock, 2015) OCR-i (*Optical Character Recognition*) ehk optilist märgi- või tekstituvastust, eesti keeles ka tärgtuvastus (nt Jaanimäe, 2021; mõlemad variandid toob välja ka EKI ühendsõnastik, 2026). See tähendab optilise tehnoloogia ja arvutitehnoloogia automatiseeritud koostoit, mille abil pildi pealt leitakse sümbolite ja tekste ning tuvastatakse nende sisu (Wang et al., 2023). OCR viitab üldiselt teksti tuvastamisele erinevatest allikatest nagu dokumentidest, kaartidelt või piltidelt ja samad põhitõed rakenduvad paljuski sõltumatult sisendist, kuid käesolevas töös räägitakse teemast tulenevalt peamiselt kaartidelt teksti tuvastamisest ja selle eripäradest.

Tekstituvastus koosneb olemuslikult kahest alamülesandest: tekstide leidmisest (*detection*) ja nende lugemisest ehk tuvastamisest (*recognition*) – leidmine tähendab, et pildi pealt leitakse alad, millel tekst esineb, toetudes kas objekti lokaliseerimisele või pildi segmenteerimisele, ning tuvastuses keskendutakse eelmises sammus leitud aladel esineva teksti dešifreerimisele (Wang et al., 2023). Alamülesanded on võimalik läbi viia eraldi või ühe leidmist ja tuvastamist ühendava (*end-to-end*) protsessina (Lin et al., 2020). Objekti lokaliseerimine tähendab teksti (või muu pildil esineva objekti) asukoha välja selgitamist ja väljundina andmist seda ümbritseva risküliku koordinaatide kaudu ja segmenteerimine selle eristamist, millised pildi pikslid kuuluvad millisele objektile pildil, andes reeglina tulemuseks riskülikust keerulisema pindobjekti (Tõnisson et al., 2019). Segmenteerimise üks rakendus on tuvastuse kergendamine seekaudu, et esiti liigitatakse kaardi elemendid klassidesse nagu kontuurjooned või kaardikirjad ja siis eemaldatakse tuvastuse seisukohast tarbetud klassid, eriti kitsaste joonobjektidega klassid nagu teed ja horisontaalid (vt nt Pouderoux et al., 2007). Leidmisel ei püüta teksti sisust aru saada. Teksti leidmine eraldi sammuna aga aitab kaasa sellele, et väljundisse satub vähem müra või muud taustainfot (Wang et al., 2023). Tavapärane variant on, et vaikumisi teeb valmis lahendus mõlemad sammud ära, kuid protsess on võimalik kasutaja poolt ka kaheks jagada ja teha leidmine ja tuvastamine eri lähenemistega. Käesolevas töös sõnaga tuvastus viidatakse tervele kahe (või enama) faasiga protsessile, kui pole teisiti toodud.

OCR-i kasutatakse igapäevaselt näiteks elektroonilistest dokumentidest sõnade ülesotsimiseks. Siiski pole see ainult viimaste aastate uurimisteema – Wang et al. (2023) toovad välja, et põhimõtte on käidud välja 1929. aastal. Ameerika ühendriikide patenti taotleti mõni aasta hiljem (Tauschek 1929/1935), ja ka uurimistöodes on OCR olnud vahendiks juba pikemalt (vt nt de Mello & Lins, 1999), kuigi varajaste tuvastusmeetodite täpsus oli nõrk (Wang, 2023).

Esimesed tärgtuvastuse algoritmid võrdlesid teksti tähti ükshaaval käsitsi joonistatud näidisega (Zhou, 2026). Taoline lähenemine andis adekvaatseid tulemusi standardiseeritud fontidega, ent selle võimekus oli vähene käsitsi kirjutatud tekstide puhul või kui tähtedes esines kvaliteedi kõrvalekaldeid (Zhou, 2026). Hilisemad OCR-meetodid lähtusid selle piiritlemisest, mis kaardi pikslid kuuluvad kokku, moodustades ühe elemendi ja mis mitte, ja selle lahendamiseks on kaks põhilist võtet (Lin et al., 2020). Ühte tuntakse nimega *Connected Component Analysis* (CCA; Närvivõrkude ja masinõppe sõnastik (kuupäev puudub) ei paku vastet, kuid käesolevas töös nimetatakse seda komponentanalüüsiks). Komponentanalüüsi eeldus on see, et piltide tekstiosad moodustuvad kõrvuti asuvatest pikslitest, millel on ühine omadus nagu värv või intensiivsus (Liu et al., 2025). Kaartide puhul leitakse selle abil piisavalt suured üksused, mis eeldatakse esindavat kaardikirju (Lenc et al., 2022). Sama eeldus kehtib loomulikult ka teiste kaardikomponentide puhul. Teine võimalus on kasutada libisevat akent. Selle puhul liigub aken ehk väiksem ala, millele mudel hetkel keskendub, üle sisendpildi ja liigitab igas akna asukohas „nähtava“ sisu kas tekstiks või mitte-tekstiks ja kõrvuti asetsevad tekstiga alad grupeeritakse järgnevas sammus kokku (Long et al., 2021). Sellise meetodi abil saab tõhusalt tuvastada erinevas suuruses ja erinevate suundadega teksti (Liu et al., 2025). Need kaks lähenemist toovad ka Wang et al. (2023) välja kui traditsioonilised tekstituvastuse meetodid.

Liu et al. (2025) selgitavad eelmainitud lähenemiste nõrkusi järgnevalt: CCA põhiprobleemiks on selle võimetuset saada järjepidevalt samahäid tulemusi piltidega, milles teksti muutujate (mh font, suurus, suund) vahelduvus on suur. Libiseva akna puhul paisub väärpositiivsete tuvastuste arv suureks, eriti juhtudel, milles tekst ja mitte-tekst on visuaalselt sarnased kirju tausta tõttu. Lisaks võib see lähenemine olla arvutuslikult kallis, kuna on vaja hinnata erinevaid akna suuruseid, katmaks kaardikirjade fondisuuruse ja suuna vahelduvust (Liu et al., 2025).

Kui kasutada OCR-i skaneeritud piltidel, milles tekst eristub taustast hästi, näiteks tekstidokumentidel, on tuvastusastmeks⁴ hinnatud isegi 99 % (Lin et al., 2020). Samas võib tuvastusaste öelda rohkem tuvastatava materjali kvaliteedi ja selguse kui mudeli võimekuse kohta (Klijn, 2008). Näiteks Eesti Kirjandusmuuseum (2024) uuendab pidevalt võrdlevat analüüsi mitmekümne eri tehisarutextituvastusmudeli täpsuse kohta. Vahed on mitukümmend protsendiühikut näiteks trükiteksti ja käsitsi kirjutatud manuskripti vahel, kuigi kõik näidised on mustvalged, sirgerealised ja ilma suurema mürata (Eesti Kirjandusmuuseum, 2024). Ükski selles võrdluses kasutatud andmestik pole aga rasterkaart. Klassikaline OCR ongi kavandatud selge struktuuriga tekstidokumentidele, milles tekst on sirgerealine (Milleville et al., 2022). Rangelt võttes ongi OCR defineeritud kui „pildifailist automaatne trükiteksti tuvastamine“ (Jaanimäe, 2021, lk 47), ja kaardid on tuvastuse seisukohast oluliselt keerulisemad kui trükitekst.

Traditsioonilistest meetoditest edasist teed sillutasid Markovi peitmudelid (*HMM, Hidden Markov Models*), nagu selgitab Zhou (2026). Nende abil saadi struktureeritud meetod tekstijadade modelleerimiseks, milles tähemärkide või nende alamosade olemasolu mingis asukohas vaadeldakse peidetud olekutena ja pildi elemente nähtavate väljunditena. Selline lähenemine arvestab, et samad nähtavad väljundid võivad müra tõttu olla erinevate tegelike peidetud olekute (tegelikult kirjas oleva teksti) tulemus ning modelleerida tuleb erinevate võimalike tekstijadade tõenäosust tekitada vastav nähtav olek ehk pildiala välimus. Andmetel

⁴ Lin et al. (2020) kirjutavad selle koha peal tuvastusastmest (*recognition rate*), mitte täpsematest mõõdikutest nagu saagis või täpsus.

sellist mudelit luues võib see õppida arvestama erinevate fontide, käekirjade ja taustmüra allikatega. Tänu Markovi peitmodelitele kasvas OCR-mudelite paindlikkus (Zhou, 2026).

Järgmine arengufaas on olnud loomulikust keskkonnast teksti tuvastamine (*Scene Text Recognition*, *STR*, eesti keeles kinnistunud tõlget seisuga 5/26 pole), millest on kaartide puhul õigem rääkida tärgtuvastuse asemel. *STR*-i puhul on tegemist tärgtuvastuse alammõistega (Chen et al., 2020). Mõistega *scene text* viidatakse loomulikus keskkonnas (Long et al., 2021) või loomulikel pildidel (Zhu et al., 2016) esinevatele tekstidele nagu sildid tänavavaates. Selle alla kuuluvad ka kaardid, sest nendes esineb lisaks tekstilisele sisule taustmüra, mille Long et al. (2021) mainivad tuvastuse väljakutsena. Võrreldes trükitekstiga iseloomustavadki loomulikus keskkonnas esinevaid tekste 1) mitte-regulaarne asukoht ja sõnede arv, 2) suuruse, fontide ja suundade suur vahelduvus, 3) võimalik piltide kehv kvaliteet (mis aga on rohkem probleem kaameraga pildistamisel) ning 4) teksti meenutavate mitte-tekstiliste elementide suur osakaal (Lin et al., 2020). Need käivad siiski piltide, mitte niivõrd kaartide kohta, kuid mõned aspektid on siiski ülekantavad.

Loomulikelt pildidelt teksti tuvastamise võtted põhinesid algselt samuti üksikute tähemärkide segmenteerimisele ja inimese poolt valitud tunnuste otsimisel ning tähemärgid ühendati sõnedeks sõnastike abil (Dörterler et al., 2025). Sellised lähenemised olid aga oma võimekuse poolest piiratud (Chen et al., 2020). Eelmainitud komponentalüüsi ja libiseva akna meetodeid on kasutatud ka loomulike piltide puhul (Naiemi et al., 2022). Loomulikel pildidel olevate tekstide puhul liiguti järjest keerulisemate andmetüüpide suunas nii teksti keerukuse kui ka piltide kvaliteedi suhtes (vt nt Naiemi et al., 2022; Baek et al., 2019) ja alates 2010ndate keskpaigast on loomulikel pildidel olevate tekstide tuvastamises normiks välja kujunenud masinõppepõhised lahendused (Dörterler et al., 2025).

Masinõpe (*machine learning*) tähendab arvuti võimelisust õppida – antud andmete alusel automaatselt ülesandeid täitma õppida nii, et programmile või algoritmile ei anta ette täpseid tegevusjuhiseid vaid ainult näiteid (Sügis et al. 2024/2025). Selle alamvaldkond on sügavõpe (*deep learning*) ehk tehisenärvivõrgud (*artificial neural networks*), mille Sügis et al. (2024/2025) seletavad lahti kui ajurakke imiteerivad tehisneuronid, mis on koondatud eraldi kihtidesse. Nimetuse aluseks ongi see, et võrgu arhitektuur on mitmekihiline ja seega ka sügav, ja närvivõrke liigitataksegi nende kihtide arvu järgi (Sügis et al. 2024/2025). Sügavõppe arenedes on hoogsalt arenenud ka masinõppemise võimalused, ja üheks keskseks uurimisalaks selles on tekstituvastus (Long et al., 2021). Sügavõppe, juhendatud masinõppe ja konvolutsiooniliste närvivõrkude (*Convolutional Neural Network*, *CNN*) abil on OCR-i tulemused viimase kümnendi jooksul oluliselt paranenud, kuid rasterkaartidega esineb endiselt raskusi, peamiselt tulenevalt tausta keerukusest ja kaardikirjade paigutusest (Milleville et al., 2022). Sügis et al. (2024/2025) selgitavad, et konvolutsiooniline närvivõrk tähendab närvivõrku, milles sarnaseid arvutusi tegevaid neuroneid on mitu, kuid igaüks nendest vaatleb alamosa kogu sisendist, näiteks mingit ala pildil. Ka sügavamate kihtide iga neuron töötleb infot vaid lähestikku paiknevate nägemisväljade neuronitel ehk alamhulgast eelneva kihi neuronitest ja nende infost. Erinevad neuronid vaatlevad erinevaid osasid sisendist, kuid otsivad neist sarnaseid mustreid, sest näiteks teatud visuaalne detail võib esineda kus iganes. Läbi mitmete kihtide koondatakse selline lokaalne info kokku keskse otsuse tegemiseks (Sügis et al. 2024/2025). Konvolutsioonilise närvivõrgu abil tehtavas tekstituvastuses saadakse pildilt korraga kätte nii teksti asukoht pikslites kui teksti sisu, automaatselt vastavalt asukohateabele (Wang et al., 2023). Seoses konvolutsiooniliste närvivõrkude kasutuselevõttuga on juhendatud

masinõpe kujunenud standardmeetodiks tekstituvastusel (Milleville et al., 2022). Üheks keskseks põhjuseks selles on, et närvivõrkudele põhinevad tekstituvastusmodelid saavad paremini hakkama taustmüra ja kaldkirjas tekstiga (Zhou, 2026). Long et al. (2021) hinnangul suurim muutus, mille sügavõpe on kaasa toonud, on see, et uurijad lähenevad ülesannetele varasemast mitmekesisemalt, erinevate nurkade alt, ja seega uurimissuuna fookus on varasemaga võrreldes vähem kitsas.

Wang et al. (2023) hindavad, et sügavõppe abil tehtavas tuvastuses saadakse paremaid tulemusi kui traditsiooniliste meetoditega. Põhjuseks toovad nad välja selle, et sügavõppe algoritmides sisaldub enamasti teksti pööramine horisontaalasendisse, siis konvolutsiooniline närvivõrk omandab pildi tekstist olulise informatsiooni ning lõpuks rekurrentne võrk rakendab semantilist informatsiooni teksti kontekstist (Wang et al., 2023). Rekurrentne võrk tähendab närvivõrku, milles moodustub vähemalt üks tagasisideahel neuronite vahelistes seostes (IT terministandardi sõnastik [EKI IT], 2026). Selline tagasisideahel võimaldab modelleerida, milline täht on sõne järgmise tähena tõenäoline, võttes arvesse juba tuvastatud tähemärke ehk konteksti. Siiski Zhou (2026) mainib, et endiselt on problemaatilised alaesindatud keeled ja tähestikud, kuid on olemas ka keelespetsiifilisi adaptoreid, tänu millele tervet mudelit ei ole vaja muuta. Olulisteks edusammudeks loevad Long et al. (2021) tuvastusel siiski töövoogu lihtsustava sügavõppe, ülesandespetsiifilised andmekogud ja algoritmid, mille abil keskendutakse näiteks uduse teksti tuvastusele, kuid mida saab siiski koondada ka üldistemateks lahendusteks, ning toetavate tehnoloogiate arengu.

Üks sügavõppega seonduv keskne teema on suured keelemudelid (*LLM – Large Language Model*) ja suured nägemisvõimega modelid *LVM (Large Vision Model)* ning nende edasiarendused *VLM (Vision-Language Model)* ehk nägemis- ja keelemudel ning mitmemodaalsed modelid (*MLLM – Multimodal Large Language Model*) (vt nt Bordes et al., 2024; Yin et al. 2024). „Suur“ tähendab nende mudelite puhul seda, et juba aasta 2023 Qwen-VL-mudelites on kokku 9,6 miljardit õppimise käigus optimeeritavat parameetrit, millest 7,7 miljardit sisaldub keelemudeli osas (Bai et al., 2023). Yin et al. (2024) sõnul on teada, et parameetrite lisamine lisab oluliselt ka mudeli võimekust. Nägemise või mitmemodaalsuse sisaldumine nende mudelite nimes viitab nende võimele lisaks tekstile sisendina vastu võtta ka pilte ja mitmemodaalsuse puhul ka teisi andmetüüpe, näiteks punktipilvi või helifaile. Selline mudel ühendab mitme alusmudeli tugevused, sest nägemismudel, näiteks tekstituvastuse mudel, võib pildi detaile mõista, kuid selle arutlus- ja järeldusvõime on nõrk ning selles osas on keelemudeli võimekus suureks abiks (Yin et al. 2024). Suurte mitmemodaalsete mudelite kasutuselevõtt on muutnud tähtsust oluliselt, sest koos mudelitega on tekkinud teksti leidmist, tuvastamist ja semantilist filtreerimist ühendav õppimise voog, ja mudelite suur treenimisandmestik võimaldab neil õppida märkama keelelisi nüansse ja konteksti abil väljenduvat informatsiooni (Zhou, 2026). Suurte keelemudelitega on seotud siiski ka nii-öelda episteemiline ebaõiglus ehk mudelite haldajate võime mõjutada seda, missugust informatsiooni mudel välja toob, ja ebaõiglust põhjustab ka keeleline alaesindatus (vt Mollema, 2025).

Mitmemodaalsete mudelite puhul üks tuvastuse tulemustes väljenduv teema on ka hallutsineerimine (vt Yin et al., 2024). Kui rääkida mudelite tuvastuse väljundite vigadest, siis mitmemodaalsete mudelite kontekstis viidatakse sellele pigem hallutsineerimisena, mis tähendab mudeli väljundeid, mis pole kooskõlas sisendiga vaid on sellele juurde „mõeldud“ või „kujutletud“ (vt Yin et al., 2024). Klassikaliste tähtsustuse mudelite puhul tegemist on lihtsalt pigem valetuvastusega – modelid võivad tuvastada tähti valesti ja võib esineda

kallutatusi näiteks treenimisandmetes harvemini esinevate tähtede puhul. Need on siiski üles ehitatud tähthaaval tuvastamisele ja ei modelleeri keelt, ehk ei mõtle välja puuduvaid sõnalõppe või korrigeeri sõna sees tähti andmaks keeleliselt loogilisemaid tuvastatud sõnesid.

Sügis et al. (2024/2025) selgitavad traditsioonilise ja masinõppele toetuva lähenemise erinevusi. Nende tekst käib tarkvaratoodete kohta, kuid on rakendatav. Tuvastuse seisukohast üks oluline aspekt traditsiooniliste ja sügavõppele põhinevate meetodite vahel on see, et traditsioonilised meetodid on deterministlikud, mis tähendab seda, et tulemus on sama andmestiku ja sama parameetrite puhul alati sama. Arenenumates meetodites tuleb sätestada nii-öelda temperatuuri nulliks, et kasvatada sama tulemuse saavutamise tõenäosust. See tähendab, et mudel annab väljundina alati kõige tõenäosema vastuse, selle asemel, et ennustatud tõenäosusjaotusest juhuslikult valida. Ka traditsiooniliste meetodite puhul võivad tulemused siiski erineda, kui sisendandmestik muutub vähesel määral, näiteks kaardilehele on kaasatud veidi laiema puhvri ulatuses naaberlehte (vt nt Sügis et al. 2024/2025). Üldiselt heaks tekstide leidmise meetodiks peetakse sellist, mis ei sõltu kirjade fondist, suurusest või suunast (Pouderoux et al., 2007).

Sõltumata rakendatud meetodist on tekstide tuvastamises olulised tulemuslikkuse mõõdikud täpsus ja saagis (*precision and recall*). Tekstituvastuses mõõdetakse sõltuvalt ülesandest nii tähemärgi kui ka sõne tasemel täpsust ja saagist. Kaardilt talunimede lugemise kontekstis mõõdab saagis seda, kui palju nimesid tuvastusmeetodiga korrektselt kätte saadakse osakaaluna kõikidest kaardil esinevatest nimedest. Täpsus mõõdab, kui suur osakaal tuvastatud sõnedest on õigesti tuvastatud (Pouderoux et al., 2007). Mohamed (2025) selgitab, et täpsus arvutatakse

$$\text{täpsus} = TP / (TP + FP),$$

milles TP tähendab õigeid positiivseid ja FP väärpositiivseid tulemusi. Saagis arvutatakse

$$\text{saagis} = TP / (TP + FN),$$

milles FN tähendab väärnegatiivseid tulemusi. On tavapärane, et lahendust ettevaatlikumaks tehes täpsus suureneb kuid saagis langeb ning tuleb leida tasakaal nende kahe väärtuse vahel. Täpsuse ja saagise hindamise tavapärane mõõt on F1-skoor. See on täpsuse (T) ja saagise (S) suhe, täpsemalt harmooniline keskmine, mis arvutatakse

$$F1 = 2 * (T * S) / (T + S).$$

Kõrge F1-skoor räägib tasakaalukast mudelist (Mohamed, 2025). Kuigi F1-skoor on kehtestatud mõõdik, selle kohta on avaldatud ka kriitikat, sest on seatud küsitavaks, kas nii erilisi mõõdikuid kui saagis ja täpsus sobib kokku panna ja kas harmooniline keskmine on kokkupanekuks parim lähenemine (Hand et al. 2021). F1-skoori jaoks pole ühist piiri, millal see on „hea“ või „halb“, vaid see sõltub lähenemisest (mudelite hindamise lävendite kohta vt Hernández-Orallo et al., 2012). Järgmises alampeatükis (1.3.2 Kaardid tuvastusmaterjalina) tuuakse välja mõningaid tulemusi mõõdiku kasutamisest loomulikel pildidel olevate tekstide tuvastamises. Käesolevas töös on mõõdik antud vahemikus 0 ja 1, ent näiteks Chiang et al. (2016) annavad selle protsentides, ja autor on toonud mõõdikud algallikate kujul. Kui mõlemad komponendid pole võrdse tähtsusega, saab hinnangut täpsustada andes osadele ette kaalud (Tõnisson et al., 2019). Andmete ja metoodika osas (ptk 3.1 Testide tulemuste hindamine) on defineeritud eeltoodud näitajate (TP, FP ja FN) tähendused selle töö raames.

Eeltoodud uurimustes on siiski keskendutud nimede tuvastamisele ja nendes puudub üks käesoleva töö eesmärk ehk erinevate aegade nimekihtide kokkuviiimine. On olemas erinevaid sõnade ja tekstide ligikaudse sarnasuse leidmise meetodeid, millest üks tuntuim on Levenshteini teisenduskaugus (Levenshtein, 1966). Selle inglisekeelsed nimetused on *Levenshtein distance* ja *edit distance*. Selle põhimõte on, et loetakse kokku, kui mitu teisendust tuleb teha saamaks ühest sõnest teine (Jaanimäe, 2021). Ühes sõnes ühe tähe muutmine annab kauguseks üks ja suurem kaugus tähendab suuremat erinevust sõnade vahel. Võimalikud muutmise viisid on tähe vahetamine, lisamine või eemaldamine. Lähenemist on illustreeritud joonisel 2. Levenshteini teisenduskaugus ei ole keelespetsiifiline. Siiski rohkelt diakriitilisi märke sisaldavate sõnade puhul võivad tulemused erineda. Näiteks Davis & de Salles (2007) on arendanud portugali keeles teisenduskauguse arvutamiseks sensitiivse teisenduskauguse meetodi, milles diakriitiliste märkidega tähed grupeeritakse vastavalt tähele (näiteks á, ã ja â), ja kui täht on sama, siis on kaugus ka null. Lisaks on olemas üldistatud teisenduskaugus (*generalized edit distance*), milles on võimalik defineerida, missugused muutused on olulisemad ja tavalisemad ning ka erinevatele muutustele erinevad kaalud (Orasmaa et al., 2010). Näiteks kohanime puhul hargtäiendid (*Väike-, Vastse-* jpt) oleksid tehnilises mõttes täiesti erinev nimi, kuna harilik teisenduskaugus on nende puhul suur. Siis on võimalik anda ette loend täienditest ja määrata nendele kaal, näiteks 0,1, saamaks teave, et tegemist on erineva kuid suures ulatuses samasuguse nimega. Üldistatud teisenduskaugus annab Orasmaa et al. (2010) hinnangul ka vähem väärpositiivseid tulemusi kui tavaline teisenduskaugus.

Mõtsatare	
Metsatare	+1 = 1
Metsanare	+1 = 2
Metsanure	+1 = 3
Metsanurg	+1 = 4
Metsanurga	+1 = 5

Joonis 2. Teisenduskauguse põhimõte.

1.3.2 Kaardid tuvastusmaterjalina

Tuvastuse seisukohast erinevad kaardid tekstidokumentidest oluliselt, sest võrreldes sirgerealise ning puhta taustaga tekstiga on sisutihedad kaardid keerulisem andmeallikas. Lisaks kaardikirjadele esitatakse nendel rohkelt muud informatsiooni leppemärkide abil. Näiteks reljeefses piirkonnas on palju samakõrgusjooni ja puudega kaetud ala võidakse esitada Eesti kehtival põhikaardil suurel arvul punktleppemärke (üksik puittaim, harvik), mis meenutavad tähti või täheosasid (Maa- ja Ruumiamet, 2025a). Kõik see tekitab kaardile tuvastuse seisukohast müra ja leppemärgid võivad ka kaardikirju kinni katta. (Seda tuleb kirjade paigutusel vältida ning nime ja joone kattumisel tuleb joon katkestada (Mõisja, 2019), kuid esineb näiteid, kus neid reegleid pole jälgitud. Teksti võib katkestada näiteks sama värviga joonistatud teejoon (Pouderoux et al., 2007)).

Lisaks kaartide infotihedusele on kaardikirjad tuvastusel omaette väljakutse juba oma iseloomu pärast. Kaardil kasutatavad märgid (kaardikirjad kaasa arvatud) kujundatakse ühtset stiili järgides (Mõisja, 2019), ent kaardikirjad pole homogeensed ehk need ei paikne sama nurga all ning on erineva suuruse, värvi ja kirjastiili- või laadiga (Milleville et al., 2022, vt ka Mõisja, 2019). Klassikaline kartograafia kasutab nimeliikide eristamiseks just eelmainitud võtteid (vt nt Mõisja, 2019) ja seega ühe kaardi kirjade pilt võib olla üsna kirju ja homogeenseid nimesid on kaartidel tavaliselt väga vähe (Chiang & Knoblock, 2015). Keerukas võib olla näiteks vooluveekogu nimi, sest nimi peab olema painutatud ehk järgima joone kuju (Mõisja, 2019). Lisaks võivad kaardikirjad olla vana fondiga, mida OCR ära ei tunne, või käsitsi kirjutatud (Martins et al., 2007), kuigi Eesti Kirjandusmuuseumi (2024) võrdlus näitab, et ilma taustmürata tekstidel saadakse nende aspektidega tänasel tasemel hästi hakkama. Mitmest osast koosneva ja seega laiali paigutatud nimede jaoks on Olson et al. (2024) käinud välja lahenduse, mis moodustab vähima ulatusega puu (*Minimum Spanning Tree*) abil potentsiaalseid paare tuvastatud sõnedest, mis asuvad üksteisele lähedal ja millel on ühiseid visuaalseid muutujaid. Lahendus võiks toimida ka sõrendatud nimede puhul. Peamised rasterkaartidega seonduvad raskused tuvastamisel tulenevadki tausta keerukusest ja kaardikirjade paigutusest (Milleville et al., 2022). Tarkvarad võivad aga tuvastada ühe nime ka mitme osana ehk tulemustes on mitu nime ümbritsevat riskülikut osalise ülekattega ning nime katkestuskohtade ei tarvitse olla reeglipärased. Näiteks Inga Schlegel (2021) on tuvastanud ajaloolisi tänavanimi ja lahendanud asjaolu riskülikute morfomeetria abil liites kokku tingimustele vastavad tulemused.

Kaartidelt tuvastuse läbiviimiseks lisavad Chiang et al. (2014) koordinaatide ümberarvutamise pildikoordinaatidest geograafilisse koordinaatsüsteemi ning andmete importimise ja eksportimise. Koordinaatidega kokkuviimisega on seotud ka asjaolu, et kuigi kartograafias püütakse järgida kaardikirjade paigutamisel objekti suhtes teatud reegleid (vt Suurna & Sisask, 2010), võib kaart olla nii infotihe, et olude sunnil tuleb nimi paigutada sinna, kus selle jaoks ruumi on, kuigi tuleb järgida, et intuiitiivne seos objekti ja kirja vahel säilib (Mõisja, 2019). Peter Jordan (2009) kirjutab, et loetavuse jaoks võib tekkida vajadus nihutada kartograafilisi sümbboleid nende tegelikust asukohast eemale, ja siis kaob nende tihe seos objektiga, kuid ruumiline seos on endiselt olemas. Seda saab üldistada ka nimedele. Automatiseeritud andmetöötlusel võib see asjaolu nime ja nimeobjekti kokkuviimist raskendada.

Waheed et al. (2025) toovad välja ühest küljest mitmemodaalsete mudelite võimekust ja teisest küljest nõrkusi pildil kujutatu asukoha määramises (geolokaliseerimine). Lisaks on teada, et mitmemodaalsetel mudelitel on endiselt raskusi ruumiliste järelduste tegemisega⁵, mida peetakse mitmemodaalsete mudelite arengu järgmiseks pudelikaelaks (Liu et al., 2026). Mainitud allikad otseselt ei käsitle pikslikoordinaatide ümberarvutamist, ent võib eeldada, et geolokaliseerimine toimub teistsuguste vihjete abil ja raskused ruumilise täpsuse mõistmisel viivad juba piksliruumis valede asukohtadeni.

Ajalooliste kaartide väljakutsete hulka lisanduvad ka kaardi kvaliteet, mis võib olla vahelduv tulenevalt kas digitaliseerimise protsessist või digitaliseeritud paberkaardi kulumisest (Chiang et al., 2016), muutunud geograafilised standardid nagu mõõtkava ebatäpsus või leppemärkide ja värvikasutuse ebatäpne või tundmatu kasutus ning kaardi tegemise aluseks olnud andmete täpsusaste, mis võib olla tänasest oluliselt erinev (Martins et al., 2007). Heaks küljeks kaartide

⁵ Näidetena küsimustest: Mis suunas asub objekt a suhtes objekt b-sse? Mis suunas pildistaja on pööranud nende kahe pildi tegemise vahel?

puhul on aga oletus konservatiivsest värvikasutusest: näiteks kohanimed on enamasti musta ja veekogude nimed sinise värviga (Mõisja, 2019; Pouderoux et al., 2007).

Traditsiooniliste meetodite puudujääkidele vaatamata on tehtud OCR-i kasutavaid uurimusi kaartidel. Taolised meetodid on aga eeldanud ohtralt eel- ja järeltöötlust (Long et al., 2021). Töötluste alla kuuluvad näiteks piltide binariseerimine ja pööramine ning tähemärkide eristamine (segmenteerimine) (Wang et al., 2023). Näiteks Lenc et al. (2022) on uurinud käsitsi joonistatud 19. sajandi Austria–Ungari katastrikaartide praeguse Tšehhi alal. Nende töövoos tekitatakse binaarkaardid kirjade leidmiseks komponentanalüüsiga ja eristatakse kirjad vastavalt sellele, kas need imiteerivad käekirja (kaldkiri, nt teede, jõgede ja mägede nimed) või mitte (administratiivüksuste nimed). Meetod koosneb kolmest sammust: tekstide leidmine (*detection*), klassifitseerimine (*classification*) ja tuvastamine (*recognition*) OCR-i abil (Lenc et al., 2022). Christian Haack (2018) on oma magistritöös kasutanud Tesseract-tööriista nimede tuvastamiseks ajaloolistelt kaartidelt, kuid tulemused näitavad tema (Haack, 2018) hinnangul, et mudel ei sobi professionaalseks kasutamiseks. Tema töö kaardid olid aga binariseerimata kujul, ja ta mainib ka ise (Haack, 2018), et kaardi tauststruktuur oli segavaks faktoriks.

Chiang et al. (2016) uurisid kaartide kvaliteedi mõju tuvastustulemustele. Neil oli kuus kaarti ja 15 testala, mille pealt koostati märgiste testandmestik. Tulemuste keskmine F1-skoor oli 46,53 %, ja see vähenes endiselt kui kaartide resolutsiooni vähendati. Tuvastusmodeliks oli Tesseract (Chiang et al., 2016). Iga kaardi resolutsioon oli siiski alla 300 dpi (Chiang et al., 2016), mida peetakse soovitatavaks piiriks tuvastuse läbiviimisel (Milleville et al., 2022). Dörterler et al. (2025) on võrrelnud erinevaid loomulikel pildidel olevate tekstide tuvastuste tulemusi muuhulgas ICDAR2013 ja ICDAR2015 andmekogumites (vt näidiseid Liu et al. 2025). Tähemärgi tasemel (*character-level*) F1-skoori vahemik on 2013 andmekogumis 0,1951 (Tesseract) kuni 0,9782 (MGP-STR) ja EasyOCR-i F1-skoor on 0,7777. 2015 andmekogumis vahemik on 0,2449 (Tesseract) kuni 0,9098 (MGP-STR) ja EasyOCR-i tähemärgi taseme F1-skoor on 0,4198. Wang et al. (2025) on kogunud ülevaateartiklisse samade andmekogumite alusel saadud tulemusi ja nende ülevaates 2013 andmekogumis nõrgim tulemus on 76,8 ja parim 95,2 ning 2015 andmetel 61,0 ja parim 90,5.

Kim et al. (2023) on loonud masinõppele põhineva töövoogu nimega mapKurator⁶ nimede tuvastamiseks suurel hulgal suuremõõtmelistel kaartidel. Nimed viiakse standardiseeritud masinloetavale kujule. Töövoog koosneb kuuest sammust: 1) kaardi lugemine ja selle jagamine väiksemateks ruutudeks, 2) tekstide leidmine ja nende ümber pindobjektide joonistamine ning tekstide tuvastamine, 3) ruutudest tuvastatud tekstide kokkuviimine terve kaardi tasandile, 4) tuvastatud tekstide võrdlemine OpenStreetMap-i (OSM) alusel loodud referentsandmestikuga teisenduskauguse abil, 5) pildikoordinaatide ümberarvutamine geograafilisse koordinaatsüsteemi ja 6) tuvastatud tekstide linkimine välise kohanimedega loendiga (näiteks OSM) võimaldamaks päringud. Nad toovad välja paljude seniste lähenemiste kitsaskohtasid, näiteks tuvastustulemuste vead, mis nende hinnangul tekitab lõhet lõppkasutajate ja tehnoloogiate vahele. Lisaks nad mainivad, et senised lähenemised nõuavad andmete edasist töötlust pärast tuvastamist ja et vajalikud võtted nagu Python-skriptimise oskus, ei pruugi olla kõikidele kasutajatele käepärased. Sellepärast nad toovad oma sihtrühmadena välja lisaks geograafidele ka ajaloolased ja raamatukoguhoidjad (Kim et al., 2023). MapKurator sarnaneb

⁶ Nimi kirjutatakse väikese tähega.

oma lähenemises antud töös läbitehtud töövoole, ent see ei toeta eesti keelt (mapKurator, 2023).

Kaardilt tuvastatud kohanime paikapanekuks saab vastavalt mapKuratori töövoole võrrelda tuvastatud sõnesid kohanime loendites (*gazetteer*) toodud nimedega (Milleville et al., 2022; Schlegel, 2021). Taoline lähenemine aga eeldab korralikku loetelu ning seda, et nimed püsivad sama koha peal, ja eriti ajalooliste kaartide puhul need tingimused enamasti ei kehti (Schlegel, 2021). Lisanduvad ka muutunud kirjaviisid, mille puhul kokkuviimiseks võib rakendada sõnade ligikaudse vastavuse leidmist, mis aga võib anda hulgaliselt väärpositiivseid tulemusi ja mis võib raskeks teha arusaama sellest, mis loendi nimega tuvastatud nimi tuleks kokku viia, ning eriti ajalooliste kaartide puhul see asjaolu, et tuvastatud nime pole loendist leida (Milleville et al., 2022). On ka teada, et morfoloogiliselt rikastes keeltes sõnastikupõhine lähenemine on raskendatud, kuna iga sõnavormi tuleb käsitleda eraldi üksusena (Jaanimäe, 2021). Kohanimede häälikulised muutused on aga vähesemad kuid mõnevõrra teistsugused (vähem reeglipärased) kui üldnimedes (Pall, 1969; ka Pall, 1977). Tüüpilisi muutusi kohanimes on kirja pildis kajastuv häälduse muutus, lühenemine, võõraks muutunud nimeosa vahetamine tuntud sõna vastu ja võõrapärase nime mugandamine (Pall, 1977).

Sõltumata valitud tuvastusmeetodist tuleb kaartide analüüsis veel lahendada sisendi ja väljundi kokkusobituvus GIS-tarkvaradega juhul, kui on soov tulemusi illustreerida, kaardiks kujundada või edasist andmetöötlust või analüüsi teha. Teemat käsitletakse põhjalikumalt peatükis 2.4.1 Andmete eeltöötlus, kuid keskne teema on see, et keelemudeliga tehtav tuvastus toimub pildi raames ehk tuvastustulemuste asukohad on antud pildi pikslikoordinaatidena, millel ei ole kokkupuudet näiteks Eesti ristkoordinaatidega.

2. Andmed ja metoodika

Töö põhiandmestikuks on praegused talunimed ning kolme erineva aja Eesti põhikaartidelt tuvastatud nimed. Nimede tuvastamiseks katsetatakse üheksat erinevat lähenemist. Põhjuslikkus testide läbiviimisel on erinev tulenevalt mudeli tulemuslikkusest. Valitud mudelid on EasyOCR, MMOCR, NewOCR, PaddleOCR ja Tesseract ning kolm suurt keelemudelit: Gemini 2.5 Pro, Gemini 3 Flash ja Qwen3-VL-32B ja lisaks on kombineeritud meetod, mis ühendab EasyOCR-i tekstide leidmise ning Gemini 3-e tuvastuse.

Töö metoodilise osa läbiviimist alustades usaldati veel OCR- ja STR-lähenemisi, mispärast esimesed katsed tehti nendega ja neid on ka töös refereeritud. Töö loomisega sama-aegselt tõusis aga mitmemodaalsete mudelite võimekus tekste tuvastada, mispärast lõplik töövoog on ehitatud nii vanema kui ka uuema mudeli peale. Mudelite valik on toimunud ühendades veebiotsingut, teaduskirjandust, tehisaru soovitusi (ChatGPT-5) ja eksperthinnanguid modelleerimaks valikut, mille ees seisab nn tavakasutaja. Töövahenditest on kasutatud Pythonit tuvastamiseks ning kaartide eel- ja tulemuste järeltöötamiseks, ning QGIS-i andmete puhastamiseks ja visualiseerimiseks.

Teoorias kirjeldati kaartidelt läbiviidava tuvastusega seonduvaid tüüpilisi probleeme (ptk 1.3.2 Kaardid tuvastusmaterjalina). Töö käigus tuli välja aga mõningaid teisi aspekte, mida käsitletakse andmete eel- ja järeltöötuste peatükkides.

2.1 Praeguste talunimede leidmine

Valmis loendit talunimedest ei ole olemas. Maa- ja Ruumiamet pakub oma WMS-teenuses Aadressiandmete süsteemi (ADS⁷) abil loodud rasterkihti „taluni“, milles on aluskaartidel esinevad talude nimed, aga see esindab ainult valikut katastriüksuste nimedest (Maa- ja Ruumiamet, 2024). Töökäigu talunimede leidmiseks pakus välja Maa-ametist⁸ Mall Leht (2025). Vastavalt sellele on antud töös ajaliste kihtide kokkuviiamise referentsandmestikuna kasutatavad praegused talunimed saadud võttes Kohanimeregistrist punktobjektid, mille tüüp on „maaüksus, krunt või talu“. Siis valiti nendest välja need, mille tunnuspunkt asub hoone sees. Selliste tunnuspunktide puhul on katastriüksuse lähiaadressiks hoone nimi, ja sellest tulenev meetodi oluline piirang ongi, et seda saab teha ainult hajaasustusaladel (Leht, 2025). Hooned võeti ETAK-i (Eesti topograafia andmekogu) andmestikust. Töö käik andis 107 822 punkti, mida saab vaadata kui nimega talu. Siiski osa nimedest sisaldab liigisõna. Liigisõna on kohanime koostisosa (KNS, 2003 § 3 lg 4), kuid mitte osa nimetuumast (KNS, 2003 § 3 lg 6). Sellele vaatamata on töös filtreeritud välja tühikuga nimed, kui need osutasid teist liiki objektidele nagu kirik, kauplus või prügil. Andmetesse jäeti objektid, mille liigisõnaks oli talu. Tulemuse läbivaatamisel selgus, et mõne taluks loetava nime juures on ka number (nt *Sireli 1*), mis mainitagu, et pole kooskõlas Kohanimenõukogu (KNN, 1995) soovitusega. Need nimed saadi kätte päringuga

"ametlik_ni" LIKE "% %" AND "ametlik_ni" NOT LIKE "%taluni" AND NOT ("ametlik_ni" ~'[0-9]').

⁷ Aastal 2026 on AKS (Aadresside ja kohanimede süsteem), ent WMS-kihi kirjelduses on endiselt vana nimekuju.

⁸ Aastal 2025 kasutusel olnud nimekuju.

Järelejäänud andmeid uurides tuli välja, et on ka rohkem kohtasid, mis on numbriga kuid pole talud, näiteks *Rohuküla sadam 8*. Liigisõnade alusel filtreeriti edasi päringuga

"ametlik_ni" LIKE '% %' AND "ametlik_ni" NOT LIKE '%talu' AND (NOT ("ametlik_ni" ~ '[0-9]') OR "ametlik_ni" LIKE '%sadam%').

Lisaks sadamatele tuvastati ainult üks liigisõnaga väljajäetav punkt (*Pallase piirkond 1a*), mis on garaažikompleks Paldiskis, ja see eemaldati käsitsi. Pole siiski välistatud, et andmetes ei saaks olla rohkem müra. Kõiki käsitsi läbi ei vaadatud ning tehisaru (ChatGPT-5) ei andnud adekvaatset vastust küsimusele, kas sisendtabelis esines veel mingi eeltoodud päringu tingimusega sarnane struktuur, milles liigisõna oleks erinev (mitte sadam).

Pärast päringute alusel eemaldamist jäi 106 800 punkti. Lisaks 14 hoone sees asus rohkem kui üks punkt (kokku 24 punkti) ja need saadi kätte arvutades punktide kogused hoonete pindobjektide sees. Juhtumeid uuriti kaldaerofotode abil, ja paistab, et kohati uus hoone on ehitatud vanasse kinni (näiteks *Nigula* ja *Uue-Nigula*), aga kindlale järeldusele jõuti ainult nende punktide suhtes, millel on ühe hoone sees sama nimi. Need esindavad püstakuid ja nende korduvad punktid eemaldati andmetest. Seda problemaatikat illustreerib joonis 3. Põhikaardil on näha, et hoone, mille peal on kuus nn talunimepunkti (antud juhul püstakut), on vare, kuid nimed on siiski olemas. Maa- ja Ruumiameti andmeid saabki tõlgendada vastavalt täpsemale uurimistemeale.

Tulebki nentida, et kasutatud metoodika ei arvesta punktidega, mis asuvad vareme peal, vaid hooned on kaasatud sõltumatult seisust. Nimesid saab vajadusel edasi filtreerida kasutades ETAK-i hoonete andmestikku, milles on tüübina toodud, kas tegemist on elu- või ühiskondliku hoone, kõrval- või tootmishoone, vundamendi või varemega, kuid kuna käesoleva töö eesmärgiks on tagasisivaade, ei peetud väljasõelumist otstarbekaks. Pärast üleliigsete punktide eemaldamist jäi 106 795 nime. Praegused talunimed esindavad nimede seisu 3/25. Edasise töö tarbeks on kasutatud nimesid, mis jäävad testalade sisse (vt hiljem tabelid 1 ja 2).



Joonis 3. Karilao 2025 põhikaart ja 2022 ortofoto koos nn talunimed punktidega. Kaartide allikas: Maa- ja Ruumiamet (2025).

2.2 Kaartide valik ja iseloomustus

Sõltumata sellest, kui kaugele minevikku vaadatakse, esimene asi kaartide valikul on antud töö kontekstis talunimed olemasolu. Lisaks sellele piiravad kaartide valikut kõige olulisemal määral kaks tegurit: kaartide kättesaadavus sobilikus failiformaadis ning nende geograafiline ulatus. Esimese põhjuse tõttu tuli piirduda Maa- ja Ruumiameti kaartidega, sest kuigi paljusid ajaloolisi kaarte on digitaliseeritud (näiteks Rahvusarhiivi ja Rahvusraamatukogu (Digar) kogud), need on saadavad ainult pildina (tavaliselt JPG-fail) ilma kohateabeta ja nende georeferentimine oleks omaette töö. Teine põhjus on nende hea katvus. Käesoleva töö ajalooline mõõde on Eesti põhikaartide ulatus taasiseseisvumisest tänaseni ehk analüüsiks on võetud järgmised kaardid:

- kehtiv mustvalge põhikaart (2026), mõõtkava 1:10 000
- vanem digitaalne põhikaart (1996–2007⁹), mõõtkava 1:10 000
- trükitud põhikaart (1994–2020), mõõtkava 1:20 000

Eesti põhikaardi loomise projektiga alustati aastal 1991 ja algusest peale on põhikaardistamise lähteandmeteks olnud välitööde materjalid ning olemasolevad kartograafilised ja statistilised andmed (Maa- ja Ruumiamet 2025b). Digitaalset põhikaarti on toodetud mõõtkavas 1:10 000 ja paberkaarti 1:20 000 ning alates 1997. aastast on terve tootmisprotsess digitaalne (Maa- ja Ruumiamet 2025b).

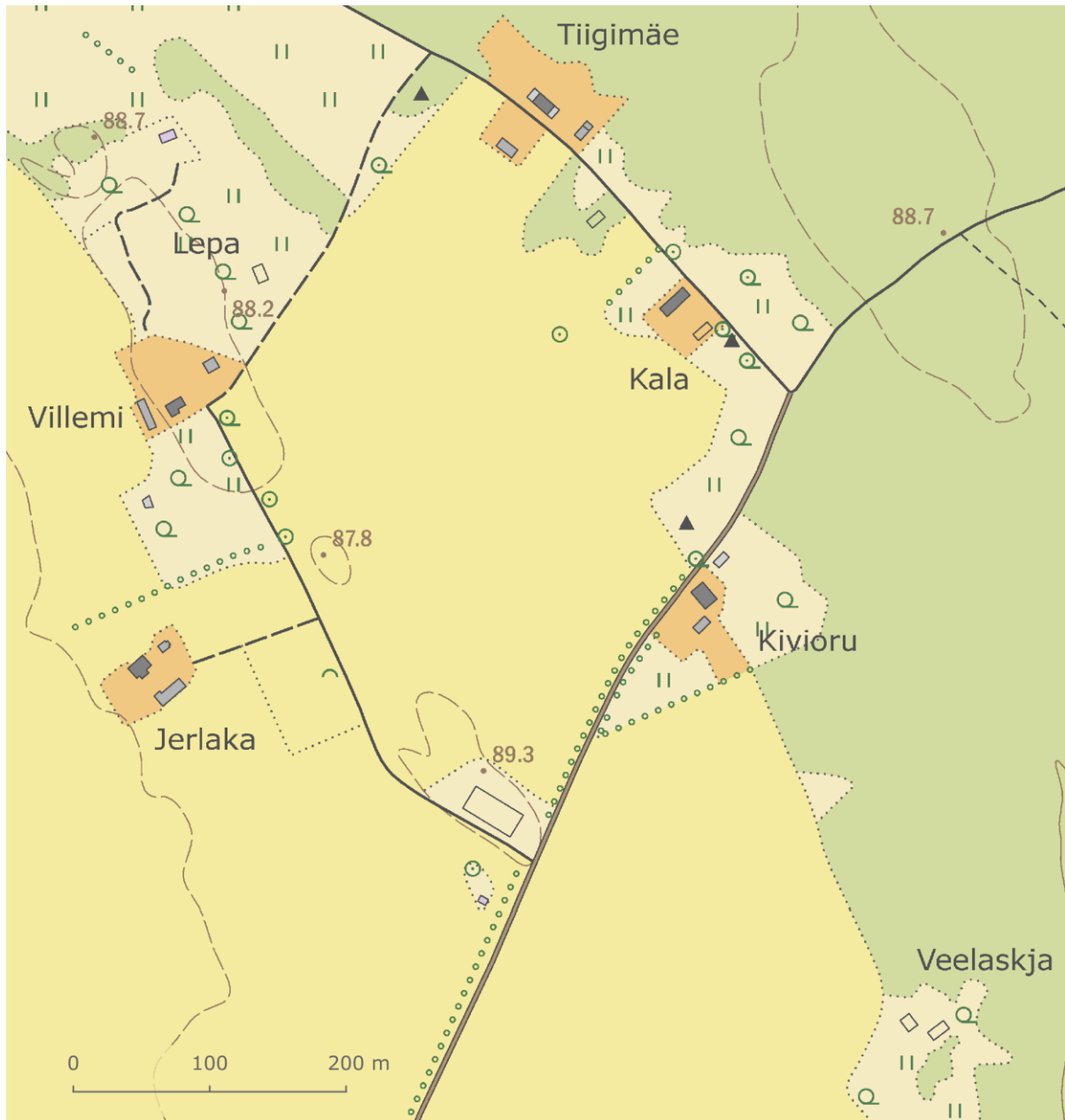
Üksikute kaartide sees esineb erinevusi, kuna kaardistamisperiood on olnud pikk. Näiteks Maa- ja Ruumiameti ajalooliste kaartide rakenduses nähtav trükikaart Iisakus on aastast 2016, Noarootsis 2004, kuid Kaereperes Raplamaal on aastaks 1994. See avaldub värvikasutuses, sest vanemad on valge taustaga ja uuemates taustvärviks on kollakas, kuid ka põhikaartide arengus on liigutud tumedatelt ja tugeva küllastusega toonidelt heledamatele toonidele. Kõikidel põhikaartidel talunimed on läbivalt horisontaalselt joondatud. Nime ja objekti vahelise seose intuiitsuse hinnangu aluseks on autori arvamus. Jooniste allkirjades on toodud kaardistuse aasta kõnealusel lehel ning asukoht vastavalt praegusele haldusjaotusele.

Talunimestik on erineva ajastu kaartidel paljudes kohtades sarnase tihedusega, kuid esineb ka erandeid. Näiteks Vaki külas Pärnumaal on talunimesid Eesti trükitud põhikaardil (1: 20 000, aasta 2016) rohkem kui kehtival põhikaardil või vanemal digitaalsel kaardil (aasta 2002). Teisest küljest Kudina külas Jõgevamaal suurim talunimed tihedus esineb kehtival põhikaardil ja kõige väiksem vanemal digitaalsel kaardil (aasta 1998). Et nimede tihedus peaks tulenevalt mõõtkavaga seotud generaliseerimise astmest olema väikseim väiksema mõõtkavaga kaardil (1:20 000 trükikaart), niisiis alati ei kehti.

Kehtival põhikaardil talunimed on musta värvi ning sama fondiga (Verdana) kui asustusüksuste nimed kuid väiksemad. Teiste rajatiste nagu kalmistute nimed on esitatud talunimedest väiksemas kirjas. Nimi on üldiselt paigutatud hoonele piisavalt lähedale, et moodustada selge seos hoonega, ning nime asukoht vastab nime paigutuse reeglitele (vt Suurna & Sisask, 2010 – talusid vaadatakse punktobjektidena, sest nimi ei mahu pinna sisse, vt Mõisja, 2019). Esineb juhtumeid, kus nimi on katkestatud joone laugest käänukohast, kuid värvilisel põhikaardil kirjad eristuvad joontest. Varemtes hoonete nimesid on kaardil ka välja toodud ja

⁹ Kõiki kaarte saab vaadata Maa- ja Ruumiameti ajalooliste kaartide rakenduses <https://xgis.maaamet.ee/xgis2/page/app/ajalooline> ning sulgudes aastaarvud on X-GIS-is antud.

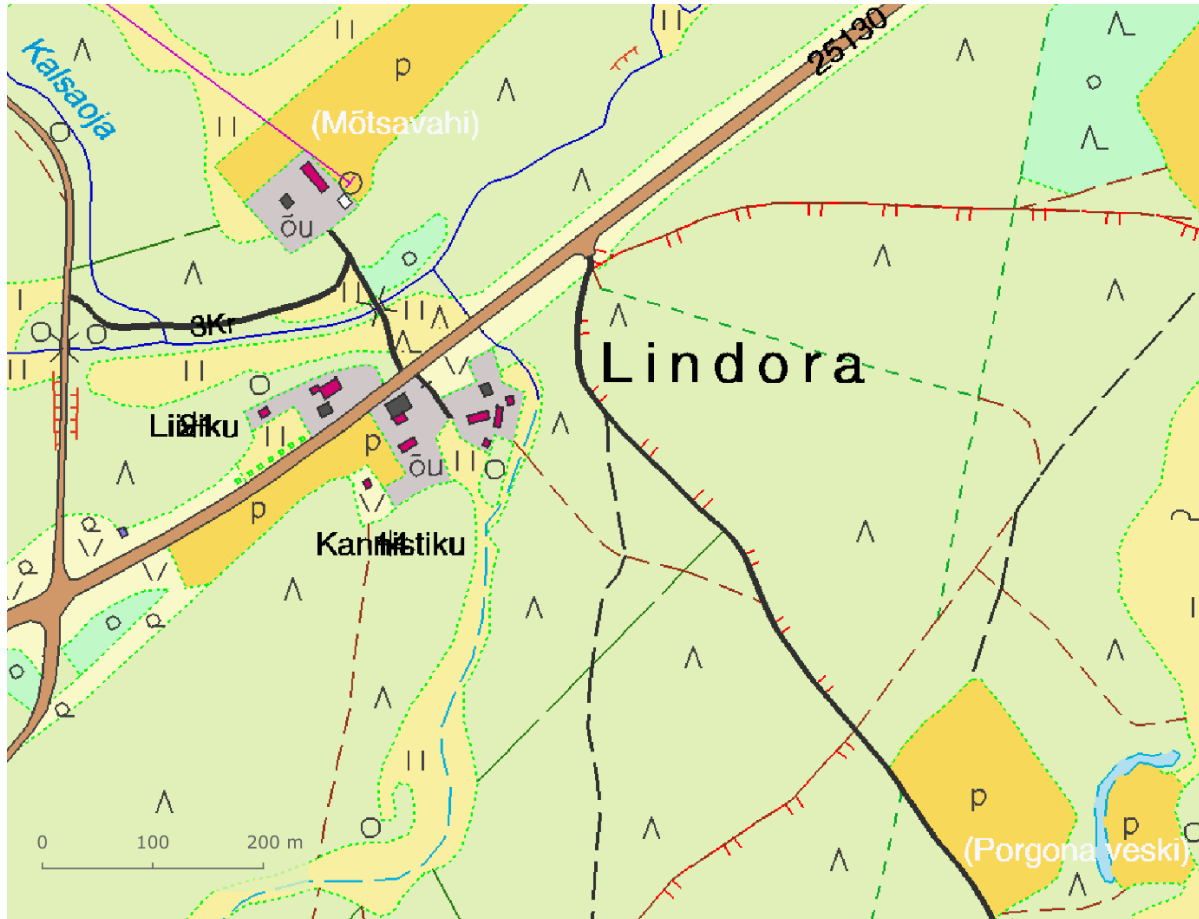
nende puhul seos võib olla kohati ebaselge, kuna vareme pindobjekt ei eristu kaardilt sama hästi kui kiri. (Mõlema eelmise kohta joonis 4.) Kehtival põhikaardil ei ole eraldi välja toodud rööpnimesid. Mustvalgel põhikaardil taustainformatsioon kohati segab nimede lugemist, sest üksikud nime tähed sulanduvad teiste kaardi objektidega kokku (hiljem töös on näidis sellisest olukorrast Koguvalt, joonis 9). Kehtiva põhikaardi lahutus on 0,25x0,25 m ja talunimede väiketähe suurus vastab umbes 12 meetrile.



Joonis 4. Nime katkestus joonega (*Lepa*) ning varemetes hoone nimi (*Veelaskja*, ka *Lepa*) kehtival põhikaardil. Võhu küla, Vinni vald. Kaardi allikas: Maa- ja Ruumiamet (2026).

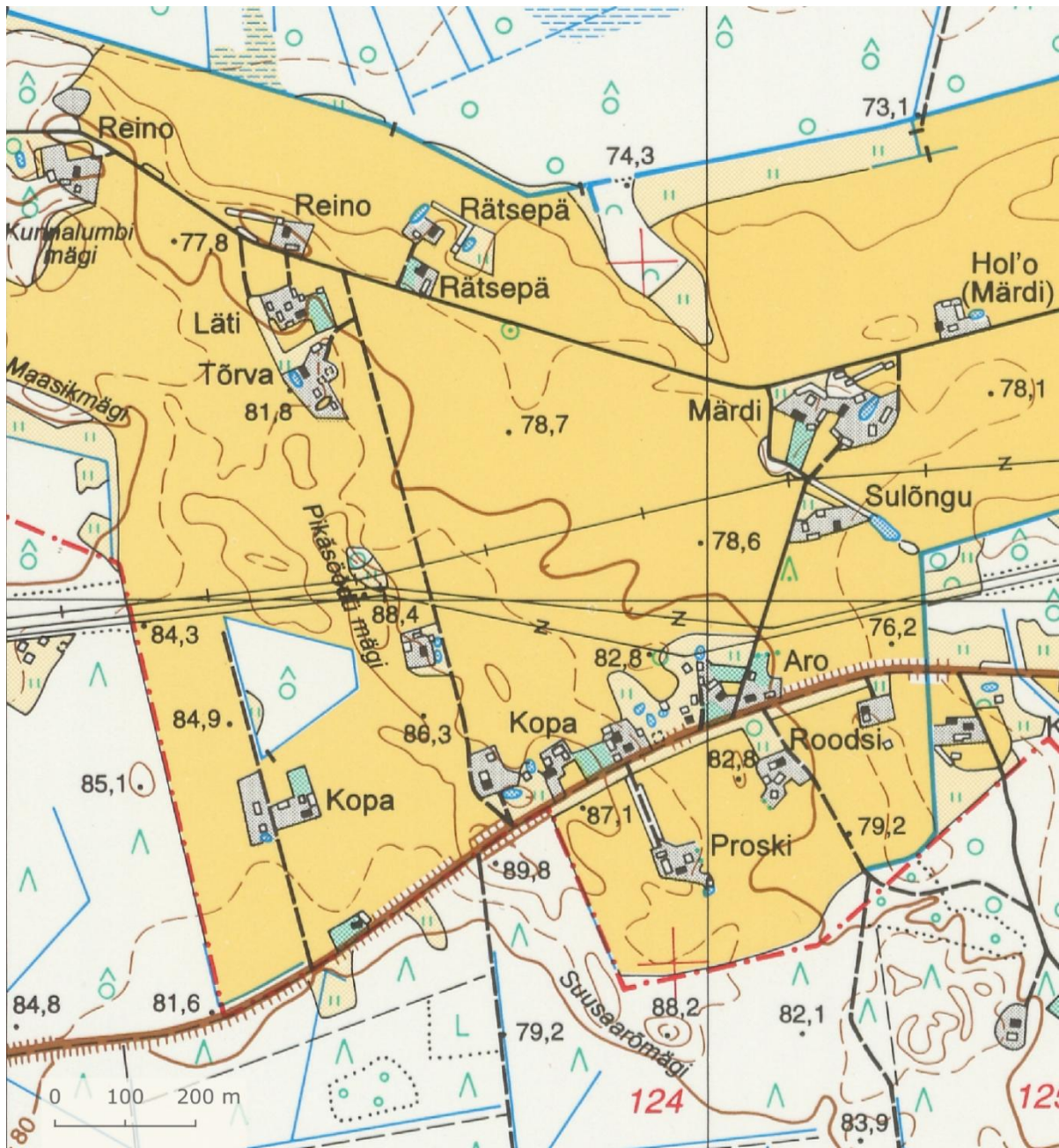
Vanemal digitaalsel põhikaardil asustusüksuste nimed on sõrendusega, talunimed on väiksema suurusega ja ilma sõrenduseta. Teiste punkt- ja väiksemate pindobjektide (nt puurkaev, park) nimed on toodud kaldkirjas. Nimede paigutust saab ka selle kaardi puhul pidada intuitiivseks ja reegleid järgivaks. Iga hoone või hoonete kogumi juurde ei ole hajaasustusaladel paigutatud nime ning kohati on märgitud ainult „õu“, kuid otsuse põhjus kaardi vaatlemisest ei selgu. Kaardil on toodud ka rööpnimesid, mis on sulgudes ja valge värviga, mispärast need ei eristu

kaardil heleda tausta vastu. Lisaks rööpnimedele on ka mõne põhinime eristuvuse puhul probleeme, sest need kattuvad kaardi punktleppemärkidega, kuid selle puhul ilmselt on tegemist veaga (mõlema kohta joonis 5). Samal joonisel on ka põllu leppemärk (p), mis kergelt tuvastatakse omaette nimeüksusena. Vanema digitaalkaardi lahutus on 0,89x0,89 m ja talunimed väiketähe suurus vastab <14 meetrile.



Joonis 5. Nimede kattuvus teiste kaardi sümbolitega ja valge värviga rööpnimed (2003). Lindora küla, Võru vald. Kaardi allikas: Maa- ja Ruumiamet (2026).

Trükikaardil suuremate asustusüksuste nimed on musta teksti ja sõrendusega, talunimed on samuti mustad, kuid väiksema suurusega ning ilma sõrenduseta. Fondiks on Verdana. Kui hajaasustusaladel on märgitud üksik hoone või hoone ja abihoonete kogum, on selle juurde üldiselt paigutatud ka talunimi. Vahemaa hoonest nimeni jääb proovimõõtmiste alusel maksimaalselt vahemikku 100 kuni 120 meetrit, kuid on intuitiivselt seostatav. Kui nime paigutus otseselt hoone kõrvale on konfliktis teiste kaardielementidega nagu koordinaadiristid, lahenduseks on enamasti olnud nime paigutamine pisut eemale, mitte selle ära jätmine, või siis nimega lõikuva joonobjekti katkestamine. Samas kõikide kohtade puhul pole kindel, kas nimi on ära jäetud ruumipuuduse tõttu või sellepärast, et tegemist polegi olnud nimega hoonega. Rööpnimed on trükikaardi vanematel lehtedel sulgudes põhinime all kuid sama musta värviga kui põhinimed, uuematel kaldkirjas põhinime all, ent esineb ka kaldkirjas nimesid ilma põhinimeta, mida saab rööpnimedeks pidada. Mõningad nimed kohati korduvad (joonis 6), mis asetab nõudmisi tuvastusprotsessis õige nime õige kohaga kokku viies. Trükikaardi lahutus on 0,84x0,84 m ja talunimed väiketähe suurus vastab <25 meetrile, mis on digitaalkaartidest umbes kahekordne, mis on loogiline seoses mõõtkavaga.



Joonis 6. Korduvad nimed trükitud põhikaardil (1996). Puusepa küla, Võru vald. Kaardi allikas: Maa- ja Ruumiamet (2026).

2.3 Testalade valik ja iseloomustus

Eesmärgiks oli valida viis testala nii, et nad oleksid võimalikult mitmekesised: et aladel leiduks erinevate keelte mõju, erineva kvaliteediga kaarte, erinevaid nimede tihedusi, erinevaid ajaloolisi kihte ning erinevaid talude paiknemise ruumilisi mustreid.

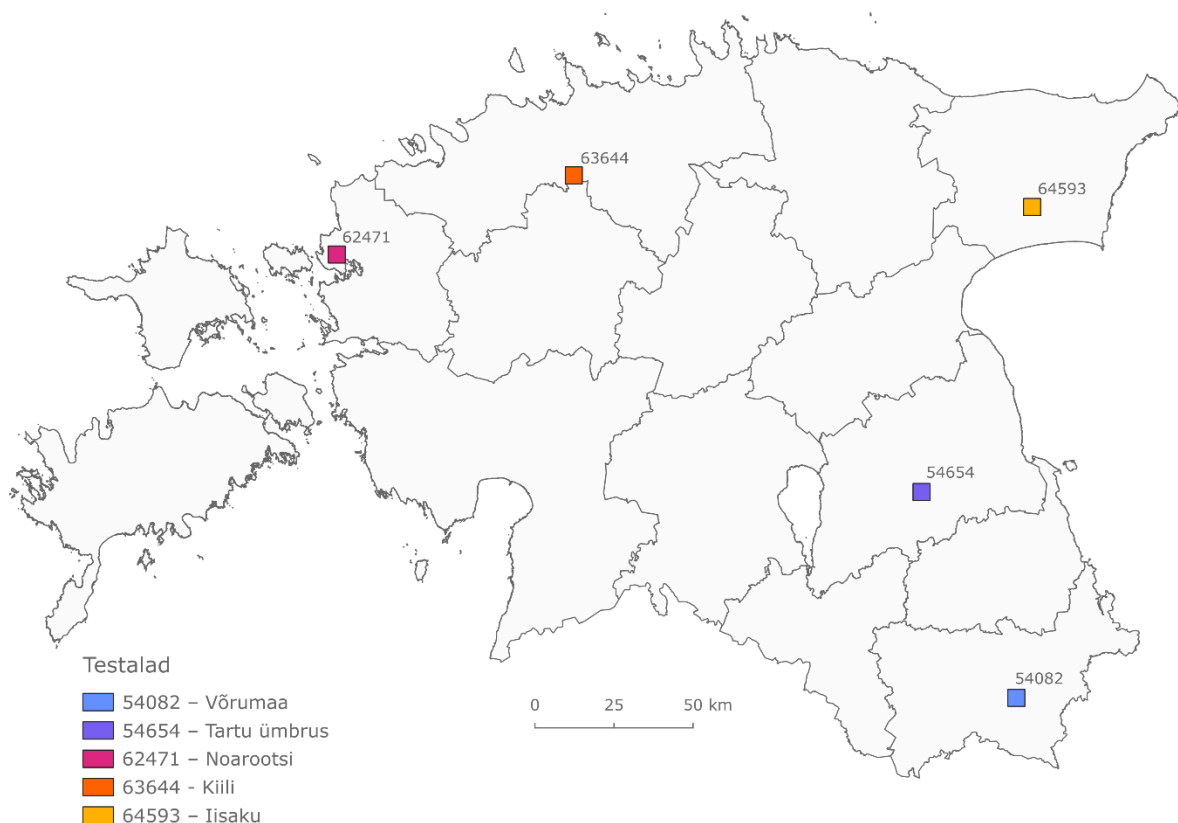
Alade valikul vaadati talunimede koguseid kõikidel analüüsitavatel kaartidel, et valimisse saaks ka nimestiku poolest mitmekülgseid kaardilehti ning et kokkuviidavaid nimesid on kindlasti igal kaardilehel ka olemas. Samuti peeti silmas ka tulevast võimalust laiendada andmestikku vanemate kaartide peale. Kõiki tingimusi polnud võimalik täita sama hästi, sest

kuigi erineva aegade põhikaartide lehed on erineva väljanägemisega, kvaliteedi erinevusi näiteks käsitsi lisatud märkuste või plekkide kujul nendel ei esine, ning külatüüpide ehk talude ruumiliste mustrite esindatust ei saadud mitmekesiseks. Laiendamise võimalusest tulenevalt kaartide kvaliteedi erinevused ongi suuremad sama alade põhikaartidest vanematel kaartidel. Ruumiline katvus asetas piiranguid ainult trükikaardi puhul, sest kümnetuhandelised kaardid katavad terve Eesti maa-ala. Trükikaardil on katteta ala Virumaal ja Koeru ümbruses. Edasise uurimise puhul tuleb aga silmas pidada, et eestiaegsete topograafiliste kaartide (eriti 1:25 000 aga ka 1:50 000) kattuvus on oluliselt väiksem, ja valitud testalad on nende pealt leida.

Valitud alad on:

- 54082 (Võrumaa)
- 54654 (Tartu ümbrus)
- 62471 (Noarootsi)
- 63644 (Kiili vald)
- 64593 (Iisaku)

Testalad vastavad kehtiva Eesti põhikaardi 1:10 000 lehe suurusele ja seega ühe ala ulatus on 5x5 km. Sellele on lisatud 200 m puhver, mille tekitamise põhjuseid selgitatakse täpsemalt peatükis 2.4.1 Andmete eeltöötlus. Kui töös viidatakse testalale, siis peetakse silmas puhverdatud Eesti põhikaardi 1:10 000 alaga sama ruumilise ulatusega ala, kuigi vanemate kaartide nomenklatuur on teistsugune. Testalade paiknemine on toodud joonisel 7.



Joonis 7. Testalade asukohad Eestis. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

Võrumaa eristub teistest oma liigendatuma pinnamoe poolest, mis kajastub kaardil suure koguse samakõrgusjoontena. Erinevat keelsust esindavad Noarootsi (näiteks *Kolanäs*, *Nygårds*), ja Võrumaa (näiteks *Kõivomäe*, *Mõtsatare*). Tartu ümbrus on arenenud hoogsalt ja

testala vanadel talumaadel asub praeguseks tehnoпарк. Kahe vanema põhikaardi vahe pole seal siiski ajaliselt suur, kuna digitaalne kaart on aastast 2005 ja trükikaart 2006. Võrumaa lehed on suurima ajalise haardega. Kiili valla testalal nimestik on tihenenud ning koos Võrumaaga on nende trükikaardid erineva tooniga. Iisaku testalale jääb samuti teise tooniga trükikaarti.

Kehtiva põhikaardi lehtedele kantud nimede ja praeguste talunimede suhtes on suuri erinevusi testalade lõikes (tabel 1).

Tabel 1. Kaardile kantud ja praeguste talunimede suhtarvud.

ALA	KAARDILE KANTUD NIMI TK	PRAEGUSEID NIMI TK	%
54082	126	132	95,5
54654	46	79	58,2
62471	83	102	81,4
63644	78	91	85,7
64593	74	85	87,0

Arvestades Võrumaa kaardilehe taustainfo tihedust on sellele mahutatud suur kogus nimesid. Tartu ümbruses omapoolt sisaldab praeguste nimede andmestik ka kõrval- või tootmishooneid – mida pole välja filtreeritud, kuna antud töös ei võeta seisukohta, kuidas vastutav instants hoone liigitanud on –, mille nime kaartidele sageli ei kanta. Vähemas ulatuses on vastav olukord Noarootsis. Kiili puhul ühest seletust välja ei joonistu, kuid Iisaku puhul on teemaks hoonestuse struktuur: Varesmetsa külas on üksteises kinni asuvaid hooneid, millest igaühel on omaette nimi (vrd püstakute teema peatükis 2.1 Praeguste talunimede leidmine). Lisaks asub servas mõni nimi, mida testandmestikusse pole kaasatud, kuna tõlgendus on olnud, et nimega seotud hoone on naaberlehe peal.

2.4 Andmete eeltöötlus, testide läbiviimine ja tulemuste järeltöötlus

Long et al. (2021) väljatoodud, nende hinnangul traditsiooniliste tuvastusmeetoditega kaasnevast paljust eel- ja järeltöötlustest pole pääsu ka käesolevas töös. Töövoog on toodud lisas 5 ja selles on esitatud ka testandmetega tehtud tulemuste kokkuviiamise osa kuid pole selgitatud praeguste nimede kättesaamist. Järgmises käsitletakse iga sammu põhjalikumalt. Väljundskriptis (lisa 4) toodud töövoog koosneb lisaks kaartide eeltöötlustele kolmest osast: tuvastusest EasyOCR-iga, tuvastusest keelemudeliga ning kombineeritud meetodist, milles tekstid leitakse EasyOCR-i abil ja tuvastus tehakse keelemudeliga. Näidiseks on toodud ka tuvastustulemuste kokkuviiimine testandmetega.

2.4.1 Andmete eeltöötlus

Erinevad kaardid nõuavad erinevaid eeltöötluste samme, kuid peamised on koordinaatsüsteemi lisamine metaandmetesse või selle ümberprojitseerimine, alale vastava pindobjekti loomine, puhverdamine, kaardilehtede lõikamine ja kokkukleepimine mosaiikkaardileheks, kaardi binariseerimine ning fakultatiivselt ka lehe ruutudeks jagamine. Lisaks töö tarbeks on tekitatud nimede tuvastuse testandmestik. Mosaiigi tekitamine on selles mõttes valikuline, et

tuvastustulemused saab kätte ka mosaiikimata lehelt, kuid mosaiigiga kaasnevad ka servas katkevad nimed. Ennem mainitud samme tuleb siiski tutvuda sellega, missugust sisendpildi tuvastusmodelid nõuavad ja vajadusel pildid ära teisendada. Töövoo koherentsuse nimel viidi eeltöötlus läbi Pythonis, kuid soovi korral saab mõned osad teha ka GIS-tarkvaras ning piltide teisendamise jaoks on olemas veebirakendusi.

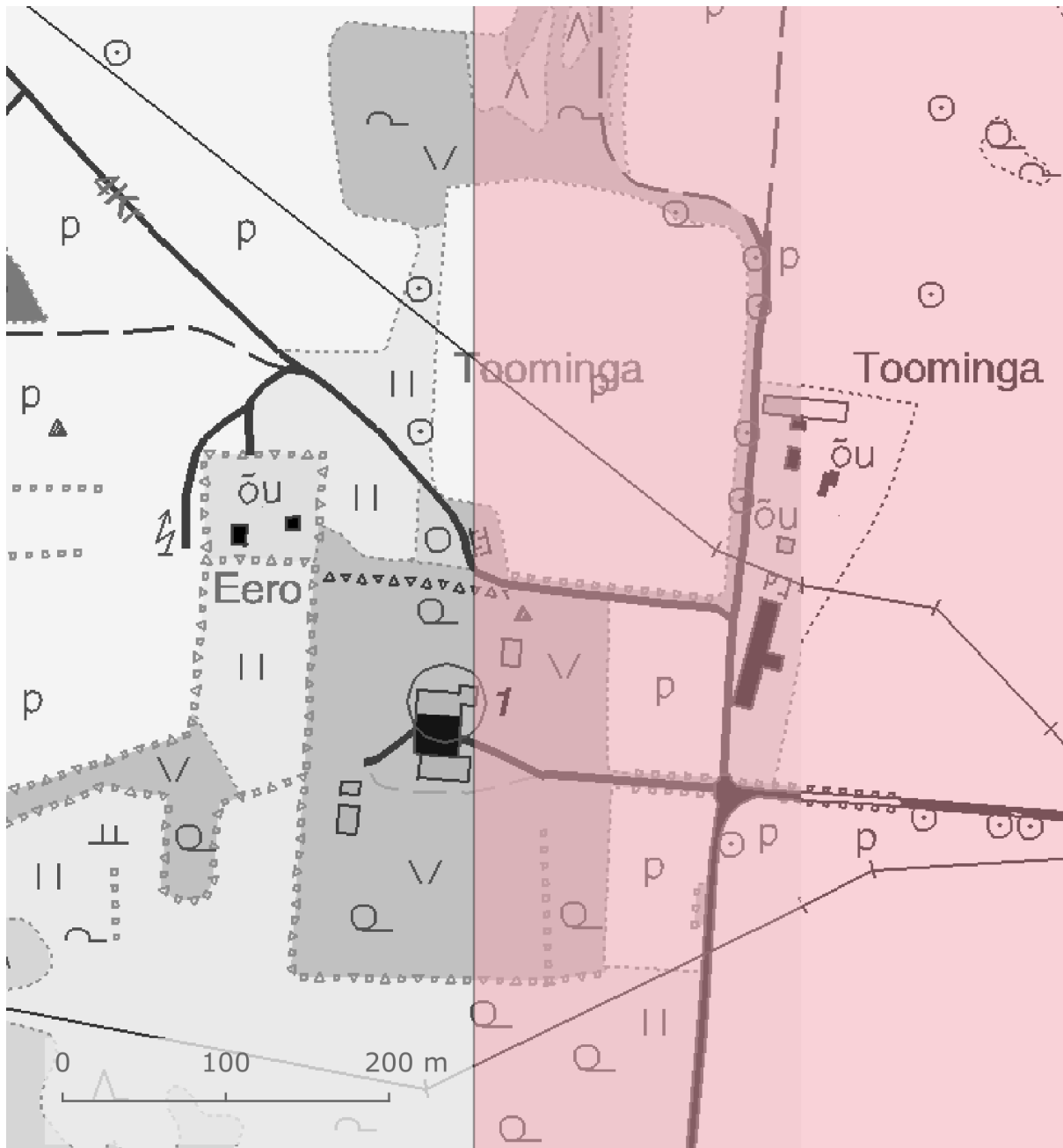
Maa- ja Ruumiameti kaardid on GeoTIFF-vormingus. Ideaalis saaks selle failitüübi säilitada, sest siis tuvastustulemusel on olemas asukoht ja koordinaatide teisendused nihkeid ei tekita. Töös katsetatud mudelitest ainult EasyOCR loeb GeoTIFF-faile ja teised nõuavad teisendust näiteks PNG-kujule. Sõltuvalt sihtfaili tüübist võib piltide teisendamine tähendada informatsiooni kadumist: teisendus tehakse kas kadudega (*lossy*) või kadudeta (*lossless*), ja JPEG-kujule viimine on kadudega teisendus kuid nii PNG kui ka TIFF säilitavad informatsiooni ka teisendamisel, mille tasuks on siiski suurem faili maht (Mane et al. 2025).. Kuigi Chiang et al. (2016) toovad välja, et pildi kokkupakkimine (*compress*) võib mõjutada kaardikirjade kvaliteeti, on antud töös valitud kokkupakkimine, kuna kaardid on hea kvaliteediga ning lisaks need on alla laaditud arvutisse.

Koordinaatsüsteem tuleb kas lisada metaandmetesse või kaardileht ümberprojitseerida. Mõlemad võtted on skripti kirjutatud. Hilisemas faasis tekitatakse pindobjekt, millega kaarte lõigatakse, ja sellepärast peavad andmed olema täpselt samas koordinaatsüsteemis, sest muidu Python formaalse keelena käsitleb neid kahe erineva süsteemina. Näiteks kehtiva põhikaardi¹⁰ mustvalgetele lehtedele on lisatud kõrgusandmed (QGIS näitab koordinaatsüsteemi kujul „Estonian Coordinate System of 1997 + Baltic 1977 height”), ja kui pindobjekt on Eesti riiklikus ristkoordinaatide süsteemis (EPSG: 3301) ilma kõrgusandmeteta, lõikamine ei õnnestu. Vanem digitaalne põhikaart aga on õiges koordinaatsüsteemis, kuid see informatsioon puudub metaandmetest („EPSG kood: None”), ja siis piisab teabe lisamisest. Lisaks keelemudelite väljund on pildi pikslikoordinaatides, mis arvutatakse algse tuvastatud kaardi abil Eesti ristkoordinaate süsteemi.

Testaladele tekitati ala hõlmav vektoristikülik, millele lisati 200 m puhver. Puhvri laiuse määrasid ära kaartidel tehtud proovimõõtmised ehk eesmärgiks oli mahutada mosaiikkaardilehe peale kõik märgised, mis viitavad testalal asuvale nimeobjektile. Mõni nimi algab digitaalsetel kaartidel väljastpoolt üksikut kaardilehte või jätkub sinna, ja nimede kokkuviimine jääb puudulikuks, kui tuvastatud saab ainult osa nimest. Seega testandmetesse pole kaasatud nimesid, mis katkevad mosaiigi servas, kui nad viitavad objektile testalast väljapool. Esineb aga ka üksikuid juhtumeid, kus pole selge, kus asub koht, millele nimega viidatakse. Näide sellisest olukorrast on toodud joonisel 8. Ajaloolisi kaarte kaasates puhverdamist põhjendavad ka kaartidel esinevad nihked, mis tulenevad nii kaartide väljavenitatusel georefereerimise tulemusel kui ka erinevate aegade erinevast kaardistuse täpsusest (vt Martins et al., 2007). Kui kasutada töövoogu ainult ühe ajaloolise kaardilehe ulatuses ja visuaalne vaatlus kinnitab, et nimed üle serva ei lähe, siis just nihete tõttu ei tuleks hoiduda lõikamise ja mosaiikkaardilehe moodustamise etapist. Asi on teine, kui soovitud tuvastusala koos puhvriga on täielikult kaardilehe poolt ära kaetud. Kasutades ainult põhikaarte ühtib nende kaardijagu selles ulatuses, et 10-tuhandeliste lehtede vähemalt üks serv on alati sama koha peal kui trükikaardil, aga erinevate seeriade ühtlustamiseks ja kõikide nimede

¹⁰ 2026 lehed. Oletusena on, et kõikidel kehtiva põhikaardi lehtedel on sama koordinaatsüsteem. Aasta 2025 lehtedel kõrgussüsteemi pole.

kaasamiseks tuleks puhver moodustada. Väljundskriptis on toodud võte puhvri tekitamiseks ühele lehele korraga.



Joonis 8. Nimi kaardi servas: vasakpoole *Toominga*, mille puhul seos kohaga pole ilmne ja seega on põhjendatud nii nime kaasamine kui ära jätmine (antud juhul kaasatud). Roosaga puhver. Vanem digitaalne põhikaart, testala 62471. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

Mosaiikkaardilehe loomiseks tuleb eelmainitud puhverdatud ristkülikuga lõigata testala kaardilehte ja seda ümbritsevaid lehti ning saadud tulemfailid kokku kleepida. Mõlema digitaalse kaardi (kehtiv põhikaart ja 1:10 000 kaart) jaoks mosaiigi loomine on lihtne, sest on saada ainult kaardialad ilma raami ja leppemärkide seletuseta ja võib otse puhverdatud alaga lõigata. Trükikaardi lehtedega aga kaasnevad raamid ja raamivälised kirjad (see ei kajastu X-GIS-teenust sirvides, sest veebirakenduses on mosaiikimine ära tehtud), ja ajaloolistel kaartidel raamid pole alati sirgedki. Lisaks näiteks eestiaegsetel topograafilistel kaartidel kaardialad ei ole ühese suurusega tulenevalt meridiaanide koonduvusest, mis teeb masinnägemise toimimise

keerulisemaks. Raamid tuleb kõigepealt lahti lõigata. Antud töös on kaardiala leidmiseks kasutatud masinnägemise teeki (OpenCV) ja loodud raami lahtilõikamiseks eraldi pindobjekt iga kaardilehe jaoks. Kasutajal tuleb lähtuda sellest, et mõningaid parameetreid võib-olla vaja muuta (funktsioonis peamiselt „tolerance“ ehk lubatud veamarginaal ning pindobjekti ulatus), eriti kui kasutada teisi kaarte, ja igati on soovitatav tulemust visuaalselt GIS-tarkvaras kontrollida. (See on aga üks koht, kus GIS-tarkvaras toimetamine eriti väiksema andmemahu puhul võib olla soovitatav, kuna risküliku joonistamine selles on lihtne, ning ajalooliste kaartide puhul võivad raamid olla hulknurksed ja servast ära katkenud. Samuti kui viga on liiga suur, on pindobjekti alati võimalik käsitsi sättida GIS-is, sest ala on kohateabefail (Shapefile, GeoJSON), mis sisaldab teabe asukoha kohta ehk GIS-is tehtud muudatused salvestatakse sellesse. Käesolevas töös on terviklikkuse huvides samm viidud läbi Pythoniga.)

Mosaiik on tuvastuse jaoks vaja binaarkujule viia, nagu on seda teinud näiteks Lenc et al. (2022) ja Pouderoux et al., (2007), et saada tuvastuse jaoks piisavalt suured kontrastid. Ainult kehtiv põhikaart on allalaaditav mustvalgena, mis nimetusele vaatamata on binaarne, ja ülejäänud lehed tuleb teha Python-skripti abil. Ka binariseerimiseks on kasutatud OpenCV-teeki ja täpsemalt selle adaptiveThreshold-funktsiooni. See eeldab, et kasutaja sätestab nii akna (*blockSize*) kui ka lävendist lahutatava konstandi suuruse. Et tegemist on paindlikku (*adaptive*) lähenemisega tähendab, et terve pildi ulatuses ei rakendata sama lävendit vaid see on erinev kaardi erinevates osades (Rosebrock, 2021). Suurem akna suurus tähendab kaardilt väiksemate sümbolite hääbumist (Rosebrock, 2021), näiteks üksikute puude, kuid tekstid on kaardi tausta vastu sedavõrd tugeva kirjalaadiga, et seal informatsiooni kadu ei olnud testalade näitel näha. (Konstandi sätestamine omapoolt avaldas testitud kaardilehtedel mõju, ja suurem tulemus andis kõrgema F1-skoori.) Optimumi otsiti tehes tuvastust erinevate väärtustega ja väärtus valiti vastavalt tuvastustulemuste kvaliteedile. Töös antud väärtused on suunised, millega võib tuvastustööd alustada, ent tuleb olla valmis väärtuseid muutma. Soovitatav on kahtluse puhul tekitada kaks erineva konstandiväärtusega kaarti ja neid visuaalselt võrrelda.

Üksik kaardileht on tuvastusmodelite jaoks enamasti liiga suur sisend ka siis, kui see on lõigatud vastavaks testala suurusele koos puhvriga. See kehtib eriti mitmemodaalsete modelitega. Sellepärast tuleb kaardileht jagada väiksemateks ruutudeks (*tiling*). Ruudud tuleb teha osalise ülekattega vältimaks nimede täielikku katkestumist. Ruutude arv sõltub kaardilehe suurusest ja tasakaaluka lahenduse leidmine on osaliselt katsetamise küsimus. Antud töös on peetud sobilikuks ala jagamist 6x6 ruudustikuks, kuid teema üle arutatakse peatükis 3.3 Metoodika hinnang ja kitsaskohad. Automatiseeritud lahendused toetavad näiteks selliseid juhtumeid, kus nimi *Servamäe* on tuvastatud kahelt kõrvuti ruudult osadena *Serv* ja *amäe* (testala 54082 vanem digitaalkaart). Siis on võimalik need koordinaatide alusel kokku viia. On aga võimalik, et ülekatte pärast tuvastatakse nimi kahe osana *Sei* ja *ervamäe*, kuna esimese puhul r-tähe serv on mudeli hinnangul „i“, teise puhul aga serv lõikab nime kaheks. Mitmemodaalsete modelite võib aga oodata oskavat teha järeldusi nime kokkuviimise suhtes.

OCR-mudelite saagise ja täpsuse hindamiseks, loodi testandmestikud iga töös uuritava mosaiikkaardilehe jaoks, sest hindamiseks on vaja erinevaid aluskaarte. Tabelis 2 on toodud talunimede kogused mosaiikkaardilehtede kaupa ehk kogused vastavad kehtiva põhikaardi 1:10 000 lehele (5x5 km), millele on tekitatud 200 m puhver. Kehtiva põhikaardi jaoks on olemas praegused talunimed (vt ptk 2.1 Praeguste talunimede leidmine), kuid tulenevalt nende tüübist (nimi punktina hoone peal märgisest eemal) ja kättesaamise meetodist (ruumipäring ilma seoseta kaardipildiga), need otse testandmestikuks ei sobi, sest see nimestik hõlmab ka

nimesid, mida kaardi peal pole, ja seega saajaprotsentiline tuvastustulemus andmebaasi väljavõttega oleks võimatu. Rööpnimed on kaasatud, kui need on kaardil välja toodud, sest need võivad moodustada olulise sideme vanema ja uuema nime või nimekuju vahel, kuigi nende roll vaheldub ja kohati nendes on säilinud vanem nimi (*Ando* ja *Palojärve*), kohati näiteks Võrumaal on sellisena toodud võõrkeelne nimekuju (*Metsa* ja *Mõtsa*), ent kohati on rööpnimi kohas, kus talu enam pole. Rööpnimesid on tuvastusel koheldud sama staatusega kui põhinimesid. Trükikaardil rööpnimed on toodud sulgudes (vt ptk 2.2 Kaartide valik ja iseloomustus), kuid kirjapanekus sulud on ära jäetud. Tuvastusmudel loeb ka need, aga need on pärast tuvastust (vt ptk 2.4.3 Tulemuste järeltöötlus) lihtsasti eemaldatavad kui nime mitesobilikud tähemärgid. Testandmestiku loomisel joonistati iga talunime ümber riskülik, sest tuvastuse tulemusel saadakse koordinaatpaar ja selle ja risküliku alusel saab olemasoleva ja tuvastatud nime kokku viia leides kohad, kus koordinaatide punkt jääb risküliku sisse. Nagu teoorias selgitatud, tuvastuse esimene samm on tekstide ehk neid ümbritsevate riskülikute leidmine, kuid mudelite hindamiseks oli vaja kindlat referentsandmestikku, kuna polnud ette teada, kui hästi erinevad mudelid leidmisega hakkama saavad ning kui täpsed leidmise tulemused on, ehk riskülikute ülekate leidmine oleks olnud lähenemisena ebakindlam. Nimed korjati värvilistelt kaartidelt, sest nende peal eristuvus on inimsilma jaoks suurem.

Testandmestikus on kokku 1084 nime. Testandmestik on koostatud käies käsitsi üle kaardilehed ja talunimedeks on loetud teatud kirjastiiliga nimed, mille sisu ei ole vastuolus talunime olemusega. Teisisõnu välja on jäetud tühikuga nimed, mille põhisõna viitab teist liiki artefaktile (näiteks *Lipniku elamu*, *Viro land*, *Noarootsi kool*).

Tabel 2. Talunimede kogused kaardilehtede kaupa.

ALA	KAART	AASTA	MÕÕTKAVA	TALUNIMESID TK
54082	kehtiv põhikaart	2026	1:10 000	126
54082	põhikaart 1996–2007	2003	1:10 000	111
54082	trükitud põhikaart	1996	1.20 000	50
54654	kehtiv põhikaart	2026	1:10 000	46
54654	põhikaart 1996–2007	2005	1:10 000	84
54654	trükitud põhikaart	2006	1.20 000	94
62471	kehtiv põhikaart	2026	1:10 000	83
62471	põhikaart 1996–2007	2001	1:10 000	69
62471	trükitud põhikaart	2004	1.20 000	46
63644	kehtiv põhikaart	2026	1:10 000	78
63644	põhikaart 1996–2007	2004	1:10 000	58
63644	trükitud põhikaart	2001	1.20 000	43
64593	kehtiv põhikaart	2026	1:10 000	74
64593	põhikaart 1996–2007	2001	1:10 000	44
64593	trükitud põhikaart	2016	1.20 000	78

Kombineeritud meetodi jaoks eeltöötluse sammuna on lisaks mosaiikkaardilehelt lõigatud vabavaralise mudeliga (EasyOCR) leitud riskülikud ja nende sisse jääv rastri osa (nimi) välja, loodud võre ja tõstetud väljalõigatud sisu sellesse nii, et iga võre silma tuleb üks nimi. Samal on loodud indekseeritud metaandmete fail, millega saab asukoha abil kokku viia algse ja tuvastatud nime. Näited nimedega võredest on toodud tulemuste osas joonistel 20 ja 21. Töös

on kasutatud 25 nimega võret (viimasesse jääb vähem, kui nimede kogus ei ole jagatav 25-ga), milles iga nimi on kõrguse ja laiusga 300 pikslit.

Eeltötluse seisukohast tuleb arvestada ka sellega, et käsurealt kasutatavatel mudelitel võivad olla oma tehnilised nõudmised. Näiteks MMOCR koosneb mitmest erinevast teegist, on tundlik nende kokkusobimisel ja nõuab mh numPy-teegist vanemat versiooni kui ülejäänud mudelid. See omapoolt eeldab teise keskkonna loomist.

2.4.2 Testide läbiviimine

Testid viidi läbi kaheksa erineva tuvastusmudeliga: viis masinõppele toetuvat mudelit EasyOCR, MMOCR, NewOCR, PaddleOCR ja Tesseract ning kolm nägemisvõimega suurt keelemudelit: Gemini 2.5 Pro, Gemini 3 Flash ja Qwen3-VL-32B. Lisaks loodi kombineeritud meetod, milles kasutatakse EasyOCR-i leitud sõnu ümbritsevaid riskülikuid ja keelemudelit tuvastamiseks ainult nende riskülikute seest. Esimene eesmärk oli aru saada erinevate tuvastusmudelite võimekusest ja võimalikest kitsaskohtadest. Esimesed viis mudelit on vabavaralised ja ülejäänud on tasulised, kuid tasuliste hinnakulu on käesoleva töö andmestiku ja koos kõikide katsetega umbes kuus eurot¹¹. Kõik testid on tehtud binariseeritud kaartidega vastavalt Lenc et al. (2022) ja Pouderoux et al. (2007) lähenemistele. Kõiki mudeleid on katsetatud vähemalt kolme erineva aja kaardi peal. Läbiviidud testid on toodud tabelis 3.

Tabel 3. Läbiviidud testid mudelite ja kaardilehtede kaupa.

Mudel →	EasyOCR	MMOCR	NewOCR	PaddleOCR	Tesseract	Gemini 2.5	Gemini 3	Qwen
Kaardileht ↓								
54082 kehtiv	X	X	X	X	X		X	
54654 kehtiv	X					X		
62471 kehtiv	X			X				
63644 kehtiv	X							X
64593 kehtiv	X	X						
54082 10T	X					X		
54654 10T	X							
62471 10T	X		X					
63644 10T	X							X
64593 10T	X			X				
54082 20T	X		X		X	X	X	X
54654 20T	X							
62471 20T	X	X						
63644 20T	X							
64593 20T	X		X		X			

Vabavaralistest mudelitest NewOCR töötab veebirakenduses, Tesseracti opereeritakse käsurealt, EasyOCR on kasutatav nii käsurealt kui ka Python-skriptiga näiteks JupyterLab-

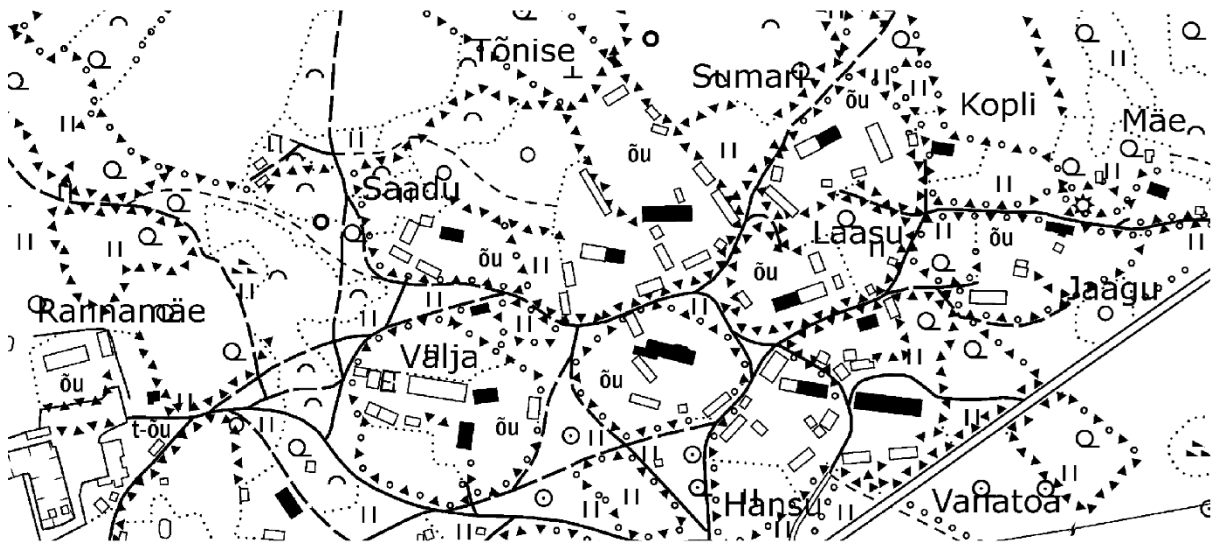
¹¹ Mudelite hinnakirjadega saab tutvuda aadressil <https://openrouter.ai/models>. „Input tokens“ tähendab viiba (prompti) ja „output tokens“ mudeli väljundi sõnade kogust. Maksumuse lõviosa oli Gemini 2.5 Pro väljundist.

keskkonnas ning ülejäänusid kasutatakse Python-skriptiga. Mitmemodaalsete mudelitega tuvastus viidi läbi OpenRouteri suurtele mudelitele ligipääsu vahendava teenuse kaudu.

MMOCR-is tuleb ise valida kasutatavad mudelid tekstide leidmiseks ja tuvastamiseks. Selles testimisel valiti leidmiseks DBNet++ ja tuvastamiseks SAR. DBNet++ on hinnatud (Bahjat, 2025) sobivat painutatud ja erinevas suuruses teksti jaoks ning SAR olevat (Li et al., 2019) samuti hea lahendus painutatud tekstide jaoks. Kuigi talunimed on horisontaalselt joondatud, leiab tuvastus erinevaid nimesid üles, enne kui tulemused ühendatakse testandmestikuga, mis aitab saada terviklikum pilt nende võimekusest kaardi peal. Täpsemalt valiti DBNetpp_resnet50-dcnv2_fpnc_1200e_icdar2015 ja SAR_resnet31_parallel-decoder_5e_st-sub_mj-sub_sa_real (vt Bahjat, 2025; Li et al., 2019).

Tasuliste mudelitele antakse ette viip (prompt). See on kui tööjuhend, milles kirjutatakse, mida tahetakse saavutada (näiteks: „loe kõik tekstid sellelt kaardilt“) ja mis kujul andmed soovitakse (näiteks: „joonista iga sõne ümber ümbritsev ristkülik ja anna selle koordinaadid xmin, xmax, ymin ja ymax“). Käesoleva töö viibad on leitavad lisadest 2 ja 3. Lisaks anti ette näidisnimi tuvastatavalt kaardilt. Mudeli tegevus on peidus ja pole teada, kas tuleks eelistada selgelt eristuvat või taustmüraga nime, kuid antud juhul valiti eristuv, et font oleks selge.

Testimist alustati kehtiva põhikaardi mustvalgete (binaarsete) lehtedega, sest need on valmis binaarkujul, mis välistab võimalikud vead binariseerimise käigus. Mõte oli, et kui mudel ei saa hea kvaliteediga kaardiga hakkama, siis on tarbetu proovida kehvema kvaliteediga lehega, mis on eriti oluline argument kaasates ajaloolisi kaarte. Iga mudeliga ei katsetatud kõiki alasid – vastavalt eelmisele argumentile kui oli kohe näha, et saagis ja/või täpsus ei ole rahuldaval tasemel, ei ole edasiminekuks abi. Nõnda saigi kaardistatud mudelite piire. Üle-eestilise lähenemise puhul näide tuvastuse seisukohast keerulisest kaardilehe osast on toodud joonisel 9, mis on Koguva külast Muhult.



Joonis 9. Näidis tuvastuse seisukohast keerulisest kaardist. Kaardi allikas: Maa- ja Ruumiamet (2026).

2.4.3 Tulemuste järeltöötlus

Pärast tuvastust järgneb järeltöötlus, mis on erinevate mudelite puhul erinev. Selles alampeatükis käsitletakse vajalikke samme üldisemalt keskendudes mudelitele, mille peale lõplik töövoog on ehitatud. Tuvastusprotsessi käigus välja tulnud probleeme käsitletakse tulemuste peatükis. NewOCR ja Tesseract annavad oma väljundi ainult tekstina, seega nende puhul tulemuste ja testandmete kokkuviiamise temaatika oleks teist laadi, ja seda järgnevas ei käsitleda. EasyOCR, MMOCR ja PaddleOCR annavad vastavalt skriptile tuvastuste asukohad tekste ümbritsevate riskülikute ja/või punktidenä. Keelemudelite väljund sõltub viiba sõnastusest ja võimalused on ka nende puhul punkt või ümbritsev riskülik. Käesoleva töö katsetel olid keelemudelid sõnade keskpunktide väljastamisel usaldusväärsemad.

Esimene samm on eelnevast tulenevalt tulemuste kirjutamine soovitud geomeetriaks. EasyOCR-i tulemuste riskülikud kirjutatakse rastrina, mis teisendatakse GeoJSON-iks. MMOCR annab riskülikute koordinaadid JSON-failina, mille saab teisendada, kuid kõik MMOCR-i nõutud teegid ei taha kokku sobitada Pythoni rastertöötluse teegi Rasterioga, mis eeldab kahte erinevat Pythoni keskkonda.

Pärast tuvastustulemuste kättesaamist tuleb neid korrastada, ja korrastamise definitsioon sõltub sellest, kas tuvastamiseks on kasutatud vabavaralist mudelit või keelemudelit. Vabavaraliste mudelite tulemused sisaldavad palju müra, eriti kirjavahemärke, mis tuleb regulaaravaldisega eemaldada. Töös on kasutatud QGIS-is järgmist avaldist:

```
trim(regexp_replace("tuvastatud_sõne", '[^\p{L}\\\\- ]', ''))
```

Väljundskriptisse on lisatud avaldise kasutamise Pythonis. See eemaldab kõik märgid, mis pole täht, sidekriips ega tühik ning eemaldab algusest ja lõpust tühikud. Kuna käesolevas töös vabavaraliste meetodite puhul tuvastuste vasted otsiti ruumipäringuga (tuvastuse punktid, mis jäävad testandmete riskülikute sisse), puhtalt arvulisi tulemusi nagu kõrgusandmeid ei eemaldatud. Keelemudelite puhul puhastamine tähendas puhtalt arvuliste tulemuste eemaldamist, nagu seda on teinud Milleville et al. (2022), ning enne seda tulemusfaili sisu korrastamist. Sisu korrastamine tähendab, et JSON-faili tulemused, mis on kujul

```
"text": "```\n\n  {\"name\": \"Tsolli\", \"x\": 377, \"y\": 115},\n  {\"name\": \"Pistümaägi\", \"x\": 622, \"y\": 120},\n"
```

(näide Gemini 2.5 tuvastusest testala 54082 vanemal digitaalkaardil) tuleb teisendada kujule

```
{\"name\": \"Tsolli\",  
  \"x\": 377.0,  
  \"y\": 115.0,  
  \"tile\": \"tile_34.png\"},  
{\"name\": \"Pistümaägi\",  
  \"x\": 622.0,  
  \"y\": 120.0,  
  \"tile\": \"tile_34.png\"}.
```

Kui soovida riskülikuid punktide asemel, tuleb valmis olla suuremaks puhastustööks, sest kõikide riskülikute nurgakoordinaadid tingimata pole samas formaadis nagu antud töö testides punktidega oli. Lisaks pole välistatud, et keelemudelid pakuvad välja enda hinnangul parimat väljundiformaati ja selle pärast ruutudeks lõikamisel on võimalik, et tulemused esitatakse erinevates ruutudes mingil teisel kujul.

Tasuliste mudelite üks omadus on lahke suhtumine kirjavigadesse, mis väljendub sellisel kujul, et üldiselt korrektse struktuuriga väljundfailis on mõningaid pisikesi eksimusi. Mudelite toimimise seisukohast need ei ole olulised, mispärast võib-olla treenimisandmetel pole sellele piisavalt tähelepanu pööratud – nagu üksikud kirjavead ei sega terve teksti mõistmist – kuid andmetöötluses tekitavad need lisasamme. Näiteks tagastavad tasulised mudelid selliseid vigadega tulemusi, kus koordinaadid puuduvad, ühe koordinaadi ümber on rohkem jutumärke kui üks või koordinaadid on valesti kirjas, kuigi viibas on selgesõnaliselt palutud tagastada x- ja y-väärtused. Sellepärast andmete puhastamise töökäiku lisati nii tulemuste normaliseerimine, kui soovitud kuju on kerge tagastada, kui ka tulemuste eemaldamine, kui soovitud kujule viimine oleks automatiseeritult liialt keeruline või täiesti võimatu. Viga võib olla parandatav ka käsitsi. Esimesel juhul on tulemustes tagastatud näiteks „y“ asemel „label_y“, teisel juhul koordinaadid ja/või nende definitsioonid puuduvad. Järgmisest näiteks on kaks erinevat tulemust:

Eemaldatakse tuvastus, millel koordinaadid puuduvad: {'name': 'Piiri', '41': 911, 'y': 911}

Eemaldatakse tuvastus, millel koordinaadid puuduvad: {'name': 'Raubamägi', 'box': [718, 189, 730, 265]}

Keelemudelitega järgmine samm on koordinaatide ümberarvutamine. Tuvastustulemuste koordinaadid on pildi pikslikoordinaadid (sisuliselt ristkoordinaadid x ja y) ja need on vaja referentsandmestiku ehk algse GeoTIFF-rastri abil ümber arvutada Eesti ristkoordinaate süsteemi. Koordinaatidega võib kaasneda viga või nihe, sest koordinaatidega tegelemine ei ole nägemisvõimega keelemudelite primaarne tugevus, ehk ka sellel sammul on soovitatav tulemust visuaalselt GIS-tarkvaras kontrollida ja punktide nihutamise suunda vaadata. Teine võimalus on mõõta nimede asukohtasid pilditötlustarkvaras ja otsida tehisaru abiga viga või nihet. Visuaalse kontrolli eelis aga on leida juhud, millal keelemudelite antud tulemused on klasterdunud iga ruudu vasakusse ülemnurka, mis võib tuleneda kas pildi suuruse vähendamisest (*resize*) keelemudeli rakendamisele eelneva kasutaja jaoks peidetud eeltötluse käigus, mistõttu keelemudel annab väljundid pildile, mida see tegelikult nägi, või keelemudeli otsusest anda väljundid suhteliste asukohtadena skaalal 0 kuni 1 (vt hiljem ptk 3.1.1 Sobivate mudelite tulemuste hindamine).

Andmete puhastamise alla läheb ka kahe osana tuvastatud nimeosade kokkusulandamine ja duplikaatide eemaldamine. Selleks kasutati testandmeid. Kokkusulandamist saab teha keeleliselt ja ruumiliselt. Antud töös EasyOCR võis anda sama testandmete risküliku sees mitu tuvastuse punkti (nime, näiteks kehtival põhikaardil testalal 54082 *Kes* ja *Tsoll* (p.o *Kesk-Tsoli*) ja *Ma* ja *Ilka* (p.o *Mallika*)) ja kombineeritud meetodi puhul oli pigem reeglipärane, et nimed katkestati sidekriipsust ja tuvastati kahe eraldi osana. Mudelite hindamiseks vabavaraliste tulemused pandi kokku vasakult paremale ning kombineeritud meetodis kui esimese osa lõpus või teise osa alguses on sidekriips ja kui riskülikute viiemeetrised puhvrid lõikuvad. Vabavaraliste mudelite puhul enesekindluseks loeti kahe tuvastuse aritmeetiline keskmine ning kombineeritud meetodi puhvri laius saadi kätte kaartidel tehtud mõõtmiste

alusel, kuid sellega tuleb leppida, et mõni nimi võib liiga pikka vahemaa pärast ühendamata jääda, kuid vahemaa tõstmine tähendaks ka mitte-kokkukuuluvate nimede ühendamist. Tasuliste mudelite olukord tulenes ruutudeks jagamisest ja ruutude servaaladel tuli ette, et üks tuvastus oli korralik ja teine oli ainult osa nimest (näiteks *epatriinu* ja *Lepatriinu*, kehtiv põhikaart, testala 54654). Nimesid ei liidetud kokku vältimaks kordumist vaid tulemustest valiti see, mis on suure tähega, ja kui see kehtis mõlema puhul, valiti pikem, sest see on tõenäolisemalt õige, kuna mudelite hallutsineerimine ei ulatu selleni, et need lisaksid tähti tuvastusele otsa. Teistsugune olukord on see, kui kohe pärast nime on kaardil sümbol, mida masinnägemine tõlgendab tähena. Töö käigus tuli välja, et ka kombineeritud meetodi hindamisel oli vaja rakendada tasuliste mudelite põhimõtet, kuna automatiseeritud indekseerimisega kaasnes vigu. Olukordades, milles ühe testandmete ristküliku sees oli mitu nime, võeti vastavalt tasulistele mudelitele eelistatult suure algustähega ja teise koha peal pikem variant arvesse. Kui mõlemad tingimused olid ühtlaselt täidetud, valiti vasakult esimene.

Saadud erinevate aegade tulemused on viimase sammuna vaja üksteisega kokku viia. Ka sellel puhul saab seda teha ruumilise läheduse alusel ühendades sellega keelelise sarnasuse. Töös on katsetatud erinevaid teisenduskauguse lävendeid ja lähenemisi ühendades tuvastustulemusi nii praeguste talunimedega kui ka uuema ajalise kihistuga. Kokkuviimise aluseks on olnud uuema kihi punktidele loodud 250 m puhver, mille seest on otsitud viis ruumiliselt lähimat kandidaati. Nendele on arvutatud teisenduskaugus ja valitud väikseima teisenduskaugusega punkt. Kui sama teisenduskaugusega kandidaate on leitud mitu, on valitud see, mille kaugus meetrites on kõige väiksem. Kõigepealt võeti praeguste talunimedede andmed ja kehtiva põhikaardi tuvastustulemused ja ühendati need, mille teisenduskaugus on 0. Siis on sama korratud ajaliselt järgmise kihiga. Lähenemise põhjuseks on see, et saada välja filtreeritud nimed, mille kuju ei ole põhikaartide ajal muutunud. Sama lähenemisega on võrreldud ka iga ajalise kihi tuvastustulemusi otse praeguste talunimedega. Lähenemiste iseloom on eksperimentaalne ja mõlemas on oma nõrkused: teine riskib vigade kuhjumisega ja esimene kõrvaldab nimede muutumise teema.

3. Tulemused ja arutelu

Optilise tekstituvastuse ehk OCR-i alla mahub palju erinevaid mudeleid. Eriti vanemad nendest on ettenähtud tuvastamiseks sirgjoonelist teksti ja eelistatult heledalt taustalt. Siiski vanema mudeli ei tarvitse tähendada halvemat tulemust, vaid palju sõltub eesmärgist ja andmestikust. (Töös kasutatud mudelid on toodud peatükis 2.4.2 Testide läbiviimine.) Tulemused on struktureeritud käsitledes esiti eestikeelsele andmestikule sobivaks osutunud mudeleid ja siis mitesobivaid. Sobivate mudelite puhul on toodud välja mudelite kitsaskohtasid ja mitesobivate puhul on selgitatud sobimatuse põhjuseid ja mõlema üle arutletud. Terve metoodika kitsaskohtasid on toodud välja eraldi alampeatükis 3.3 Metoodika hinnang ja kitsaskohad. Tulenevalt töö eesmärkidest tulemusteks on mudeli sobivus või mitesobivus, mitesobivuse põhjused, kuna probleemid ei olnud ette teada, ning saagise ja täpsuse mõõdik F1-skoor. Arutelu on lõimitud alampeatükkidesse. Joonistel testandmestiku riskülikud on musta servajoonega ja tädiseta, tuvastusmudelite joonistatud riskülikud on täitevärviga. Kõiki teste ei viidud läbi sama andmetega – näiteks NewOCR-i puhul kasutati ekraanitõmmiseid ja mitmemodaalsete mudelite puhul ruutudeks lõigatud GeoTIFF-faile, mõlemad PNG-kujul. Mitesobivate mudelite tulemuste tagasihoidlik kvaliteet aga annab mõista, et ühtne andmestik poleks päästnud. Tulemustes tuuakse tavapärasest rohkem välja näidiseid selle kohta, mis ei toiminud näitamaks, kus jookseb mudelite võimekuse piir ning põhjendamaks seda, miks on liigutud edasi teiste mudelite peale.

Väljundina esitatud töövoog on ehitatud eraldi EasyOCR-i ja mitmemodaalse mudeli peale, sest lihtsama kaardi jaoks üks nendest võib olla piisav. Lisaks on loodud kombineeritud töövoog, milles EasyOCR-iga leitakse nimed ehk neid ümbritsevad riskülikud ja tuvastus viiakse läbi keelemudeli abil. Keelemudeliks valiti Gemini 3, sest selle tulemustes esines pisut vähem kirjavigu kui Qweni tulemustes, kuigi vahe ei olnud eriti suur, ning see on arenenum kui Gemini 2.5, kuid mudelit on lihtne vahetada.

3.1 Testide tulemuste hindamine

Eesmärgiks on mõista, mis mudel sobib missuguse kaardi analüüsiks. Eesmärgile lähenetakse täpsuse ja saagise (*precision and recall*) harmoonilise keskmise ehk F1-skoori kaudu. F1-skoorid on esitatud lisades 6a ja 6b ja järgmistes alampeatükkides kirjeldatakse neid laiemalt ning selgitatakse iga mudeli jaoks sobiliku kasutusvaldkonna ulatust. Skoor on arvutatud ainult sobivate mudelite jaoks, sest mitesobivate puhul oli näha, et täpsus ja saagis on töövooleloomiseks liiga madalad ja et automatiseeritud tekstitöötlus ei suuda väljundit viia kasulikule kujule. Arvutamiseks on leitud tuvastatud punktid, mis lõikuvad testandmestiku joonistatud riskülikutega. Vasteta nimesid on käsitletud NULL-idenä.

Skooritabelites on esitatud täpsus tulemuste töötlemata kujul – välja arvatud juhul, kus ühe testandmete risküliku sees on mitu tuvastatud nime (punkti), siis on nimed kokku pandud nagu selgitatud peatükis 2.4.3 Tulemuste järeltöötlus – ja töödeldud kujul, milles nimedest on eemaldatud kirjavahemärgid nagu ' ja ~. Hindamiseks tuvastatud nimede sarnasust testandmetega arvutati igale tulemusele Levenshteini kaugus, mille annab näiteks QGIS-i sisseehitatud valem *levenshtein("tuvastatud_sõne", "õige_nimi")*.

Tulemustes on defineeritud õige positiivne (TP), väärpositiivne (FP) ja väärnegatiivne (FN) järgmiselt:

TP – õigesti tuvastatud nimi, mis jääb testandmestiku risküliku sisse ja mille teisenduskaugus pärast puhastamist korrektse nimega võrreldes on 0 (või 1)

FP – tuvastatud nimi, mis jääb testandmestiku risküliku sisse kuid mille teisenduskaugus pärast puhastamist korrektse nimega võrreldes on >0 (või >1)

FN – kaardil esinev nimi, mida kas ei leitud või mille teisenduskaugus ületab lubatud piiri.

Definitsioonid lähtuvad tuvastuse täpsusest, mitte sõnade leidmisest.

Kuna keelemudelid tagastavad pikslikoordinaadid, mitte Eesti koordinaatsüsteemi ristkoordinaate või geograafilisi koordinaate, ja ka pikslikoordinaadid on hinnang, nende ümberarvutamise tulemuste täpsus on vahelduv. Sellepärast tabelis toodud tasuliste mudelite saagis on tegelikust nõrgem, kuna kõik tuvastatud punktid pole sattunud testandmestiku riskülikute sisse. EasyOCR-i sõnade asukoha tulemused on suurema usaldusväärsusega.

Tuvastuse väljundgeomeetria on vabavaraliste mudelite puhul punktid ja vähimad ümbritsevad riskülikud, tasuliste puhul ainult punktid. Tasuliste mudelite puhul testiti risküliku moodustamist, aga see ei tulnud rahuldaval tasemel välja. Vabavaraliste mudelite andmetabelitesse on lisaks nimele kirjatud tuvastuse enesekindlus, tasuliste mudelite puhul seda tagastada ei palutud.

Vanema digitaalse põhikaardi näitajaid tõmbavad alla rööpnimed, mis on kaardil valge värviga (vt varem töös joonis 5) Tekstid ei eristu hästi ka värvilise kaardi peal ja binariseerimise protsessis kaovad need ära. (Nimi jääb sinise taustvärvi puhul.) Samuti seda F1-skoori, mis näitab puhastamata tulemusi teisenduskaugusega 0, tõmbavad samal kaardil alla sulgudes rööpnimed. Mudelid tagastavad ka sulud, kuigi testandmestikku neid pole kaasatud.

3.1.1 Sobivate mudelite tulemuste hindamine

Sobivateks mudeliteks osutusid EasyOCR, mitmemodaalsed mudelid ning eelmiseid ühendav kombineeritud meetod. Lähenemisi esitletakse antud järjekorras.

EasyOCR-i eelis on ruumiline täpsus – tuvastustulemused asuvad nimede kohal. Nõrkuseks on valed tähed ja kirjavahemärgid tulemustes ning paiguti teksti (seda ümbritseva risküliku) puudulik leidmine. Teisest küljest saab EasyOCR hakkama ka mittehorisontaalsete ja kergelt painutatud nimedega, mis tuli välja testala 54654 kehtiva põhikaardi mosaiikkaardilehe tänavanimede puhul, mida vaadati osana tulemuste kontrolli enne ruumipäringuga filtreerimist. Tulemused on nõrgemad kaartidel, millel on palju taustainfot, ent väga head selgete kaartide peal. EasyOCR on üsna jäik kuid see-eest väga lihtne ja toetab paljusid keeli, mh ka eesti keelt.

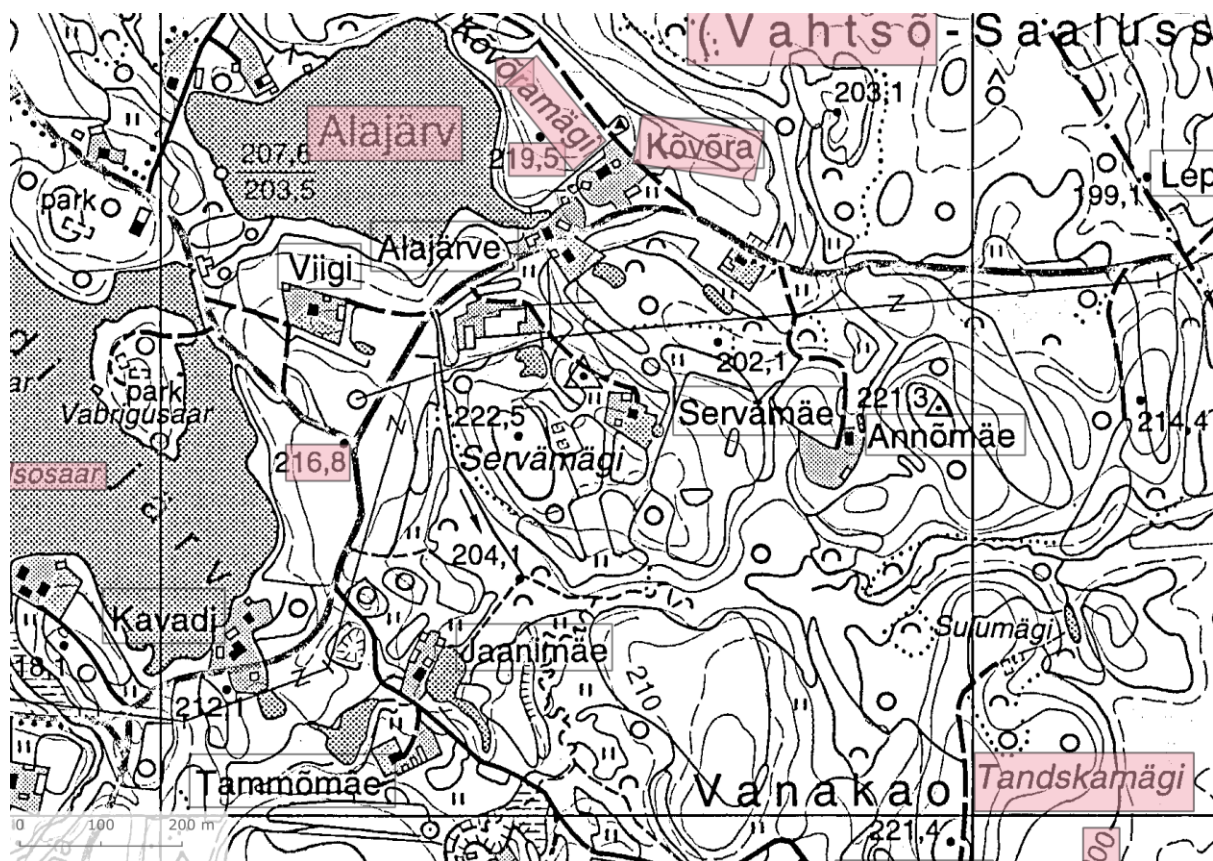
EasyOCR-iga tuvastati iga mosaiikkaardileht, aga järgnevas keskendutakse võrdlema kahe testala (54082 ja 62471) kehtiva põhikaardi tulemusi. Noarootsi poolsaar ja Võrumaa kuplistik on reljeefi poolest väga erinevad, mis kajastub ka leppemärkide tiheduses. Võrumaa mosaiikkaardilehel on tihedalt samakõrgusjooni ning kuplite vahel väikeseid põllulappe ja märgalasid, millel on leppemärk. Seega nimede taustal esineb müra, mis mõjutab tulemuse täpsust ja saagist. Noarootsis omapoolt on lagedaid alasid ja nimed paljuski valge tausta vastu.

Testala 54082 (Võrumaa) mosaiikkaardilehel EasyOCR leidis vaste 102-le nimele ehk $102/126=80,95\%$. Arvutades Levenshteini kaugust, on täpseid vasteid puhastatud kujul 52 tk 102-st ehk $50,98\%$ ja koos sellistega, mille kaugus on üks (1), 75 tk 102-st ehk kokku $73,52\%$. EasyOCR tagastab lahkelt kirjavahemärke tulemustesse – eriti erinevaid pindu piiravad punktiirjooned tõlgendatakse tihtipeale koolonina (*såvisaare:*) ja sama näide osutab ka lahke „å“ kasutuse. Kirjavahemärgid on lihtne ära puhastada, mis on EasyOCR-i tulemustega ka tehtud, ja ka mitte-eesti tähtede jaoks saab kasutada lisaks selles töös toodule regulaaravaldisele näiteks sensitiivset teisenduskaugust (vt Davis & de Salles, 2007), kui tahta tuvastus läbi viia EasyOCR-iga. Suurem teema on siiski teksti puudulik leidmine: sõne algusest või lõpust jääb tihtipeale puudu. Näiteks *Võro-Ando* asemel on tuvastatud ainult *ndo* või *Kõivsaare* asemel *vsä?*.

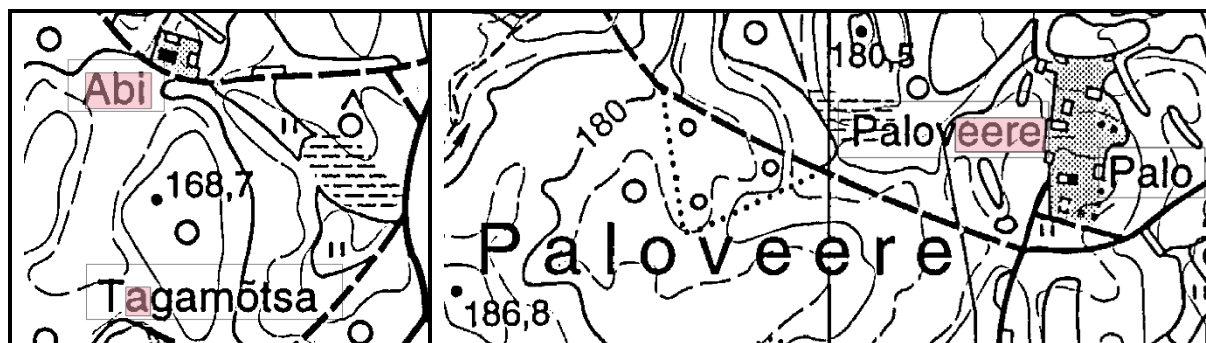
Noarootsi (testala 62471) kehtiva põhikaardi mosaiikkaardilehel on talunimesid 83 tk. Nendest 83-st vaste leiti 76-le ehk $91,56\%$. Arvutades Levenshteini kaugust, on täpseid vasteid 56 tk 76-st ehk $73,68\%$ ja koos sellistega, mille kaugus on üks (1), 70 tk 76-st ehk $92,10\%$. Ühte suurema kaugusega nime seletab ära risküliku puudulik ulatus (*Sinika* asemel *Sin*), kuid teiste puhul ühest hüpoteesi pole.

Eriti nõrk tuvastustulemus on Võrumaa trükikaardi peal (joonis 10). Teejooned eristuvad tumedamatena kui tekstid, mille võiks arvata segavat tuvastust, kuid töö andmete alusel süstemaatilist seost ei paista. Taustainfo seos tuvastustulemusega polegi üheselt selge. Joonistel 11a ja 11b on toodud kahe lähestikku asuva nime tuvastus. Joonisel 11a *Abi* on korrektselt tuvastatud kuid *Tagamõtsa* Levenshteini kaugus on kaheksa, kuna tuvastatud on üksainus täht. Kaart on aga mõlema nime ümber oma taustainfo poolest sarnane ja tume teejoongi on lähemal korrektselt tuvastatud nime kui mittekorrektset. Sarnaselt seletamatu olukord on joonisel 11b. Kui ainult tumedam toon oleks seletuseks, siis peaks suuremas kirjas toodud külanimest olema tuvastatud vähemalt osa. Talunime *Paloveere* algus kaob taustainfo sisse, mis saab olla põhjuseks selle puudulikule tuvastusele, aga inimsilma jaoks *Palo* eristub taustast võrdväärset. Kuna tehisarüütsused on oma iseloomult must kast – mille üle diskuteerib põhjalikult näiteks Bathaee¹² (2018) – saab võimalikke põhjuste üle küll arutleda, kuid kindla seost esitada ei saa.

¹² Kuigi Bathaee (2018) kirjutab masinõppest ja tehisarust, tegelikult iga piisavalt keeruline süsteem on kasutaja jaoks must kast. Näiteks kasutaja jaoks pole selge, millise taustmüra eemaldamine või mis muu võtte aitaks tuvastust tõhusamaks teha. EasyOCR kasutab ka tehisinärvivõrku.



Joonis 10. Väljalõik Võrumaa trükikaardi tuvastustulemustest. Aluskaardi allikas: Maa- ja Ruumiamet (2026).



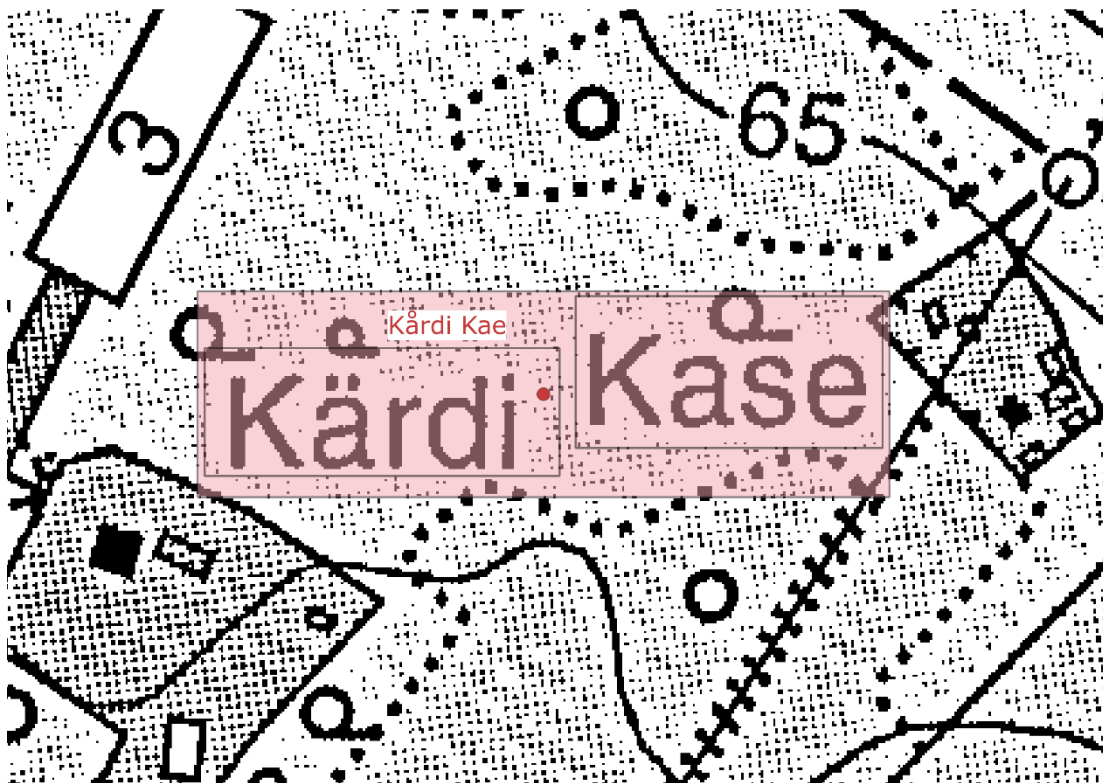
Joonis 11a-b. Puudulikult leitud nimed a) *Tagamõtsa* ja b) *Palo*. Trükikaart, testala 54082. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

Kui võrrelda käesoleva töö puhastamata andmeid Dörterler et al. (2025) loomulikel piltidel olevate tekstide tulemustega (nad ei maini, et tulemusi oleks puhastatud), on kehtival põhikaardil EasyOCR-i F1-skoor Noarootsis samas suurusjärgus kui võrdluse parima EasyOCR-i (0,7 vs. 0,7777, ICDAR 2013 andmekogum). Võrumaa skoor 0,46 jääb sellest alla ja kõikide Dörterler et al. (2025) testitud mudelite võrdluses asetub ka allpoole, kuid teiste mudelitega kõrvutamine ei anna kõige paremat võrdlusmomenti.

EasyOCR talunimede tuvastustulemusi ei mõjutanud kehtiva põhikaardi puhul testala 62471 näitel see, kas tuvastuskeelteks valiti eesti ja rootsi või ainult eesti. Ainus vahe oli ühe mitte-talunime lõpus „a“ või „ä“. Samas rootsi keelele põhinevad nimed on vanemal digitaalkaardil eestipärasemal kujul (*Nygorsi*, kehtival põhikaardil *Nygårds*) või rootsi mõju avaldub ainult kaashäälikuühendites, mida eesti keeles ei esine (*Grentsi*, rootsi keeles gräns='piir'), kuid mis

tuvastuse seisukohast ei oma erilist tähtsust. Samuti ei täheldatud tuvastuse raskusi, mis oleksid olnud seotud võrukeelsete nimedega, vaid nõrgem saagis oli pigem seotud märgiste halvema eristuvusega. Küsimusele aga ei lähenetud kvantitatiivselt, kuna piir murdelise ja üldkeelse sõna vahel võib olla hägus ja see küsimus jääb töö fookusest välja. Üks tähelepanek oli ka see, et kuigi EasyOCR eeldab tuvastuskeele valikut ja toetab eesti keelt, esines tulemustes korduvalt tähti, mida eesti keeles ei kasutata (nt *Sõramäe*, *Kõpliaia* ja *Sõerl* (p.o *Sõeru*), näidised erinevatelt kehtiva põhikaardi lehtedelt). Nende osakaalu kõikidest tähtedest ei arvatud, kuid see on pigem väike. Kuna mudeli sisse ei näe, ei saa öelda, mis nähtust põhjustab, aga oleks huvitav teada saada, mis kaaluga keele valik on.

Mõnel juhul EasyOCR tuvastas kaks kõrvuti asuvat nime ühena, sest nime riskülik on leitud ühena (joonis 12). Inimsilma jaoks kartograafia spetsiifiline lahendus on paigutada teine nimi pisut kõrgemale. Siis ebaühtlane joondus annab mõista, et tegemist on erinevate nimedega. Statistikas on *Kärdi* tulemuseks pandud *Kärdi Kae* ja *Kase* jäetud vasteta, kuna aluseks on punkti asukoht testandmestiku risküliku sees. Mõningad nimed jäid vasteta ka sellepärast, et tuvastuspunkt sattus kahe lähestikku asuva nime testandmete riskülikute vahele.

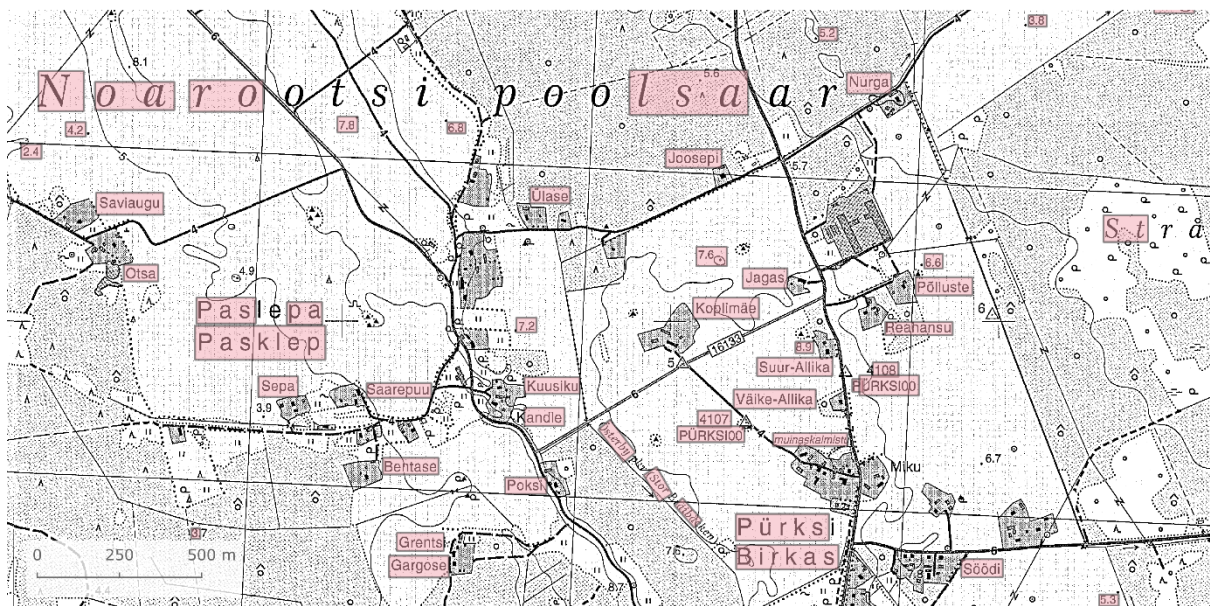


Joonis 12. Kaks kõrvuti asuvat nime (roosaga EasyOCR-i väljundgeomeetria ja punasega tuvastatud nimi *Kärdi Kae*). Trükikaart, testala 54654. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

Trükikaardi puhul katsetati EasyOCR-i jaoks kahel testalal (54082 ja 54654) kahte erinevat väärtust binariseerimise lävendist lahutatavale konstandile. Võrumaa kaardil vahed väljenduvad kohtades, kus trükirastri väärtus on madal. Tartu ümbruses erineva väärtusega kaartide vahed on inimsilma jaoks märgatavamad, mida võivad mõjutada ühest küljest Tartu ümbruse lagedad alad oletatavasti skaneerimisest tuleneva säbrulise struktuuriga ja teisest küljest Võrumaa tihedad samakõrgusjooned. Suurema konstandiga tulemus on heledam, mis aitab tekstide eristuvusele kaasa ja seekaudu tõstab nii täpsust kui ka saagist. Ainus reeglipärasus, mis autori hinnangul täpsuses eristub, on see, et suurema konstandi puhul on

vähem loetud teisi kaardi sümboteid nime osaks, kuid tähelepaneku kohta statistikat pole arvutatud ning kirjavahemärgid eemaldatakse puhastamise käigus. Sellele vaatamata on Võrumaa tulemused mõlema mõõdiku seisukohast nõrgad, kuigi täpsus on parem kui saagis.

Talunimed on Eesti põhikaartidel sellest seisukohast tänuväärseid tuvastatavad, et need on kirjutatud sõrenduseta. Joonisel 13 on toodud näidis EasyOCR tuvastustulemusest trükikaardil testalal 62471. Joonisel üks riskülik esindab ühte tuvastustulemust ja on näha, et suure sõrendusega nimed nagu *Noarootsi poolsaar* ja *Paslepa* tuvastatakse mitme osana, mis tähendab, et näiteks nimes *Noarootsi* tuvastustulemus võib olla ka ainult üksiktäht. Teistel aladel samasuure sõrendusega nimesid ei esinenud, ja kuna Noarootsis tasulisi mudeleid ei katsetatud, ei saa anda hinnangut, kas põhjuseks on mudeli valik.



Joonis 13. Sõrenduse mõju tuvastustulemustele. Trükikaart, testala 62471. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

EasyOCR-i tüüpiline viga on nimede puudulik leidmine: sõne algus või lõpp puudub või hargtäiendiga nimi on tuvastatud kahe eraldi osana. Järgmine on lihtsam edasise töötlemise seisukohast, sest osad saab kokku sulandada. Riskülikute läbivaatamisel leiti 54 vigast pindobjekti – võttes arvesse kõik eeltoodud vea tüübid –, mis moodustab testandmestiku 1084 nimest 4,98 %. Korralikke hulka on arvatud ka mittehorisontaalsed riskülikud, kui nimi on täielikult risküliku sees.

EasyOCR-i tuvastustulemus peegeldab kaardi taustainfo vaesust. Ülejäänud kehtiva põhikaardi lehed on oma kaardipildilt Võrumaa ja Noarootsi vahepealsed ja ka tulemuste (lisa 6a ja lisa 6b) F1-skoor seda näitab. Skoorid on pigem siiski kõrgema poole kaldu. Testide alusel saab öelda, et EasyOCR annab ainsa võttena adekvaatse tulemuse lehtedel, millel ei esine palju taustmüra, kui arvata välja vanema digitaalkaardi valge värviga nimed. Tuvastamata jäänud nimedes ei nähtud tendentsi lisaks kaardi mürale. Keerulisemate kaartide puhul on soovitatav valida arenenumaid mudeleid ja teha tuvastuse töövoog kombineeritult kahe erineva mudeliga kaasates teiseks tasulise mudeli.

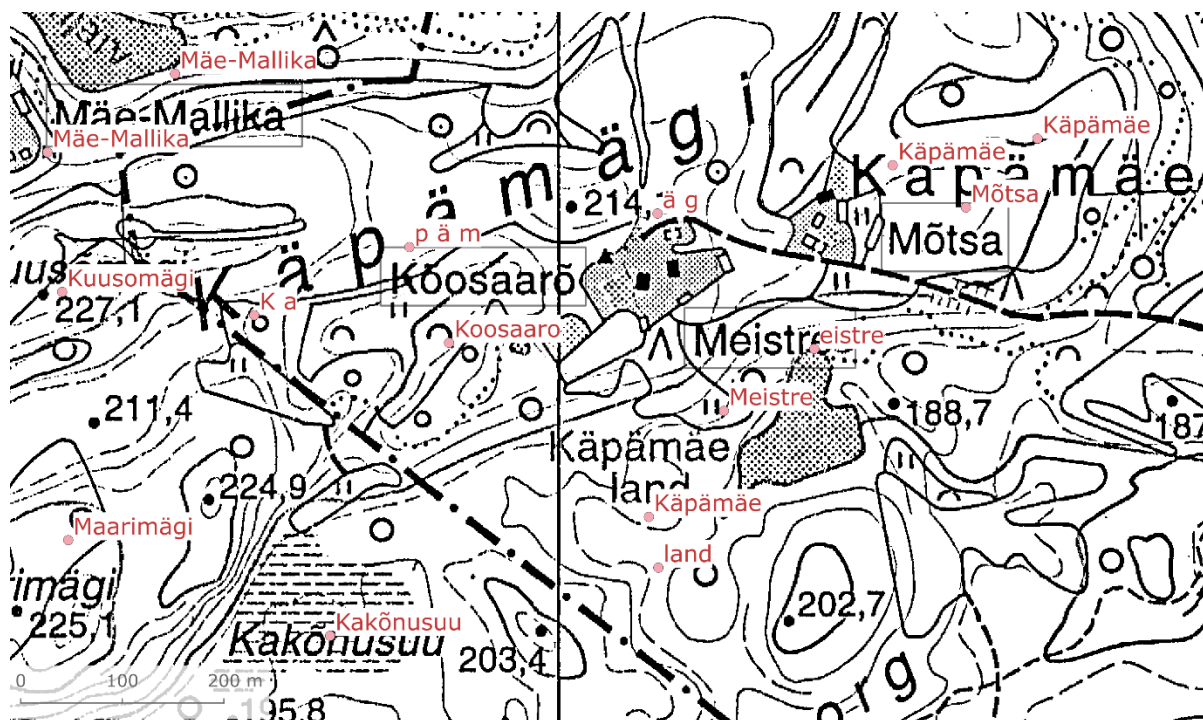
Nagu teooriaosas mainitud, mitmemoodaalsete mudelite tugevus on võime täheldada keelelisi nüansse ja konteksti abil väljenduvat informatsiooni (Zhou, 2026) kuid nendel on raskusi ruumilise täpsusega. Sellest järeldub, et ühendades tuvastustulemusi testandmestikuga jääb

saagis tegelikult vähesemaks kuid täpsus on väga heal tasemel. Qwen tuvastas natuke rohkem täpitähti või võõrapäraseid tähti kohtades, kuhu need ei kuulu, kui Gemini mudelid, kuid puudub alus öelda, kas nähtus on süstemaatiline või juhuslik. Arvutades kõikide EasyOCR-i ja mitmemodaalsete mudelite F1-skoori aritmeetilise keskmise, on see praktiliselt sama, kui teisenduskaugus on 0–1 (0,79 vs. EasyOCR 0,80), kuid teisenduskaugusega 0 mitmemodaalsete mudelite keskmine on veidi kõrgem (0,70 vs. 0,63).

Iga mitmemodaalse mudeliga katsetati vähemalt ühe Võrumaa (testala 54082) kaardilehega, mis tulenevalt taustainfo tihedusest osutusid EasyOCR-i jaoks keerulisteks. Lihtsustamaks tuvastust lõigati üks hästi eristuv nimi kaardi pealt välja ja anti mudelitele näidiseks. Ainult Võrumaa trükikaardi peal viidi testid läbi kõikide kolme mudeliga. Selle näitel Qwen tuvastab Gemini mudelitest natuke rohkem täpitähti või võõrapäraseid tähti kohtades, kuhu need ei kuulu (*Päloveere*, *Jaaniikiwi*), kuid tegemist võib olla ka erandiga, kuna teistelt lehtedelt on leida ainult *Möldri* (vanem digitaalne kaart, testala 63644 – sama ala kehtival põhikaardil siiski tuvastatud õigesti) ja *Kassõru*. Võrukeelsed elemendid nimedes ei tekitanud süstemaatilisi raskusi – vigu esines nagu *Savisaar* (ilma lõpu õ-ta, Qwen) ja *Koivomäe* (Gemini 2.5, mõlemad trükikaardi pealt), kuid need olid erandliku iseloomuga ja võivad olla põhjustatud ka muust. Võrreldes EasyOCR-iga on keelemudelite täpsus eriti palju parem Võrumaa trükikaardi peal. Lisaks kaasneb nendega vähem müra ehk teisi kaardisümboleid nagu punktiirjooni või pikseldunud joonte osasid. Keelemudeli võib arvata oma treenimisandmete alusel mõistvat, missugused sümbolid sõnades esinevad. Qwen jääb tulemustes enamasti Gemini mudelist alla nii Võrumaa trükikaardi kui kehtivate põhikaartide puhul, kuigi järgmiste puhul Gemini 3-ga tuvastati Võrumaad ja Qweniga vähema taustainfoga Kiili kaarti (testala 63644). Qwen-mudelid on treenitud inglise ja hiina keele alusel (Bai et al., 2023), mis võib põhjustada saadud tulemust. Gemini kohta treenimisandmete keeli polnud leida. F1-skooridest siiski ei maksa järeldada, et mudeli valik oleks kõige määravam tegur.

Mitmemodaalsete mudelite murekoht kaartidelt nimede tuvastamisel on nimelt see, et nad teevad arvutused puhtalt pikslikoordinaatides, sest neid ole kavandatud asukohtadega arvestamiseks, ja ka need pikslikoordinaadid on hinnang. See tähendab viga koordinaatides, mis tuleb iga kord näidiste alusel või visuaalse vaatluse abil üles otsida ja ümberarvutamisel arvesse võtta, kuid antud töö näidistel viga pole niivõrd süstemaatiline vaid pigem juhuslikuma iseloomuga, ja tulemuslikkus võib sõltuda ka näidiste valikust. Ka süstemaatiline tähendab antud kontekstis pigem seda, et ühe tehtega saadakse enamus tulemustest paika, ja viga on hakatud otsima alles siis, kui on olnud näha, et punktide asukohad ei ühti testandmetega peaaegu mitte üldse. Arvutamisel katsetati erinevaid vigade suurusi ja suundasid ja valiti see, mis andis parima saagise, kuid antud töös ei suudetud leida valemit, millega kõik tuvastatud nimed saaks testandmestiku riskülikute sisse. Sellest järeldub, et keelemudelite täpsus on järjest nõrgem kui saagis, sest hea täpsusega nimesid jääb arvesse võtmata. Joonisel 14 on toodud näide Gemini 2.5 tuvastusest Võrumaa trükikaardi lehel: osa punktidest jääb testandmestiku pindobjekti sisse, ent mõnda tuleks liigutada pigem põhja ja mõnda pigem lõuna suunas (joonise faasis duplikaate pole veel eemaldatud). Jooniselt on näha ka, et näiteks *Meistre* testandmestiku sisse jääb puudulikult tuvastatud nimevariant. Ümberarvutamisel on juhtunud ka seda, et mõni tuvastuspunkt sattub nime all oleva rööpnime risküliku peale, mis pärast tuvastuse ja kaardil kirjas oleva nime vaheline teisenduskaugus on tegelikult suurem. Mitmemodaalsete mudelite täpsus on siiski niivõrd heal tasemel, et koordinaatide ümberarvutamise vaev on põhjendatud, kui tahta kasutada ainult mitmemodaalset mudelit. Siiski probleemiks jääb, et kõik tuvastuse punktid ei saa korrektselt paika. Asjaolu saab

vähemalt kehtiva põhikaardi nimikihistu puhul parandada viies tulemusi praeguste nimedega kokku, kuid seda pole töövoos toodud, ja koordinaatide arvutamisel on selle töö näitel tasuliste mudelite puhul alati viga sees.



Joonis 14. Gemini 2.5 tuvastus Võrumaal. Trükikaart, testala 54082. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

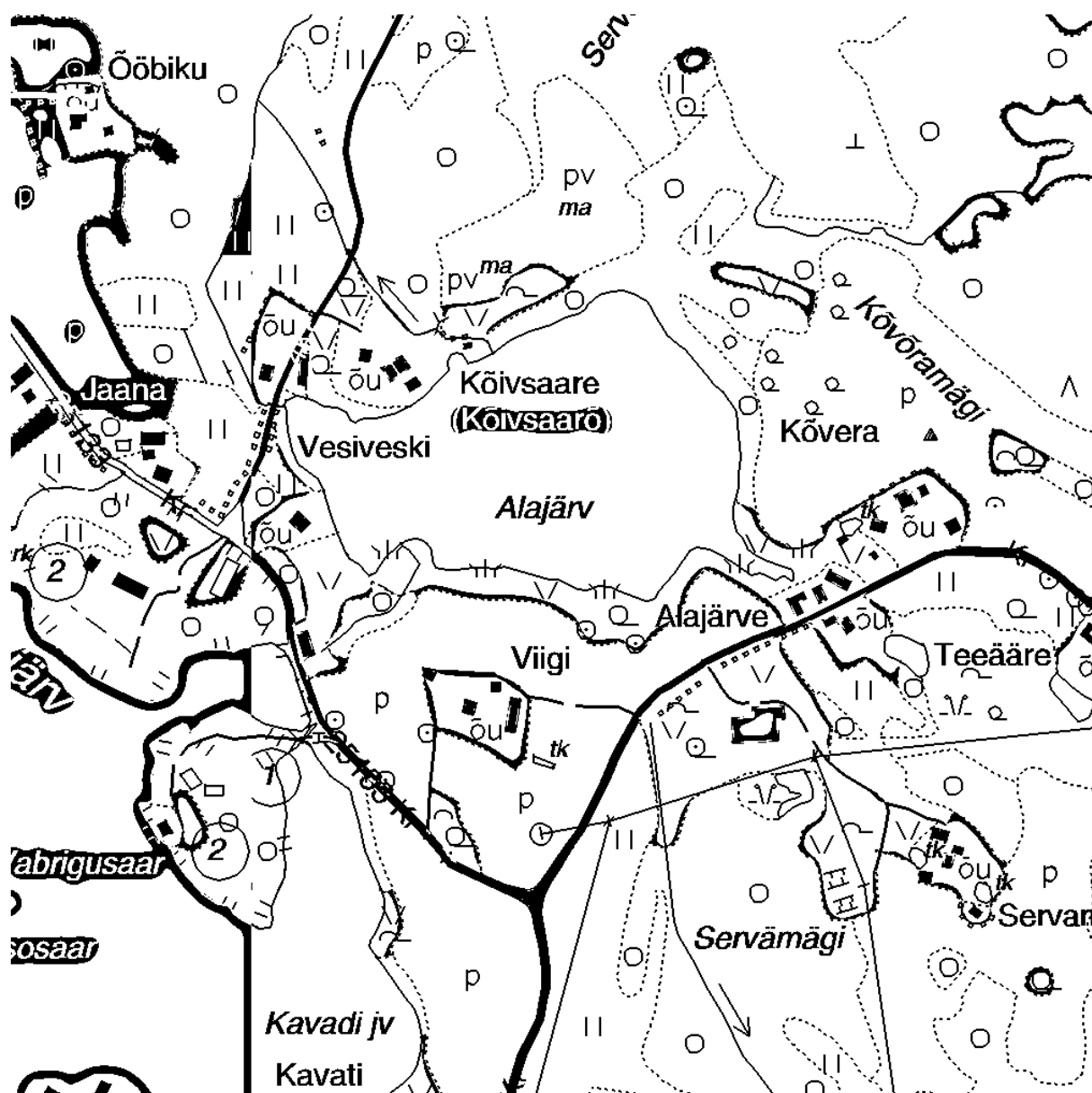
Tehisarude musta kasti probleemiga ei saa teada, mis mudeli arvutamisel täpsemalt toimub, aga testimisel oli täheldada, et Gemini 3 andis süstemaatilisema koordinaatide vea kui Gemini 2.5, mispärast selle saagis käesoleva töö testides oli natuke parem. Siiski tuleb meele pidada, et kuigi tuvastusskriptides on sätitud temperatuur nulli peale, see ainult tõstab seda tõenäosust, et tulemused on koherentselt samad. Lisaks tehisarude abil tehtaval vea otsimisel tuleb mõõtmised kas teha nime ülemnurgast – mis aga ei tarvitse haakuda keelemudeli nime ümber joonistatud kasutajale nähtamatu riskülikuga – või teada anda, mis kohast nime väärtus mõõdeti, sest mudeli koordinaadid algavad vastavalt konventsioonile ülevalt vasakult. Lahendamiseks võib minna vaja liitmistehet, korrutustehet ja/või offset-väärtust. Kuna keelemudel pole joonistanud pindobjekti välja, võib-olla loogilisem mõõta väärtus teksti keskelt, mis omapoolt mõjutab arvutuse valemit. Siiski tuvastuse saagisest ei saa otse järeldada, et tuvastus iseenesest poleks olnud edukas, vaid probleem on koordinaatides.

Sama koordinaatide problemaatika väljendub ka selles, et kohati tulemuste punktid klasterduvad iga tuvastusruudu vasakusse ülemnurka (joonis 15). Peatükis 2.4.3 on toodud kaks hüpoteesi selle põhjuste kohta, kuid kindel tõde on autorile teadmata. Selle lahendamiseks on vaja lahutus- ja korrutustehetest keerukamat lähenemist, ent ka selle töö raames pakutud lahendus tingimata pole piisav.



Joonis 15. Punktide klasterdumise näide Gemini 2.5 tulemustel. Kehtiv põhikaart, testala 54654. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

Kuna mudeli sisse ei näe, vastuseta jääb ka küsimus, miks mõnda nime ei tuvastata. Näiteks testala 54082 vanema digitaalkaardi peal on kaks tuvastuseta ruutu (puhastamata toorandmed ehk põhjuseks pole koordinaatide ümberarvutamise viga). Ühel on ainult üks nimi ja ka see pole horisontaalne, kuid teisel on nimesid rohkem ja mõni nendest ka selgelt eristuv valge tausta peal. Järgmine ruut on toodud joonisel 16.



Joonis 16. Tuvastuseta ruut. Vanem digitaalkaart, testala 54082. Kaardi allikas: Maa- ja Ruumiamet (2026).

Kombineeritud meetodi tulemused on paremad kui eeltoodud eraldi meetodite, kuna selles ühendatakse mõlema tugevused: EasyOCR käib hästi ümber geomeetriaga ja mitmemodaalne mudel tuvastab nimed korrektselt. Suurimaks nõrkuseks ongi, et EasyOCR kohati leiab nimed puudulikult. Kuna kombineeritud meetod toetub EasyOCR-i leitud ristikülikutele, selle saagis ei saa olla suurem kui EasyOCR-i oma. Selle eeliseks EasyOCR-i tekstide lugemise vastu on siiski keelemudeli võimekus.

Kuigi EasyOCR-i ristikülikutega lõigatakse välja kõik leitud nimed, hinnang käsitleb ainult talunimesid. Nad mahuvad paremini võre silmade sisse kui pikad loodusnimed, kuna kiri on väiksem ja sõrenduseta.

Kombineeritud meetod on enamasti parem kui üksik mudel, ent mitte alati, kuigi täpsuses ollakse EasyOCR-ist ees. Tulemus teisenduskaugusega 0 või 1 võib olla sama EasyOCR-il ja kombineeritud meetodil nagu on Noarootsi (testala 62471) vanema digitaalkaardi puhul: üks nimi on jäänud EasyOCR-il leidmata. Puhastamata (teisenduskaugusega 0) on kombineeritud

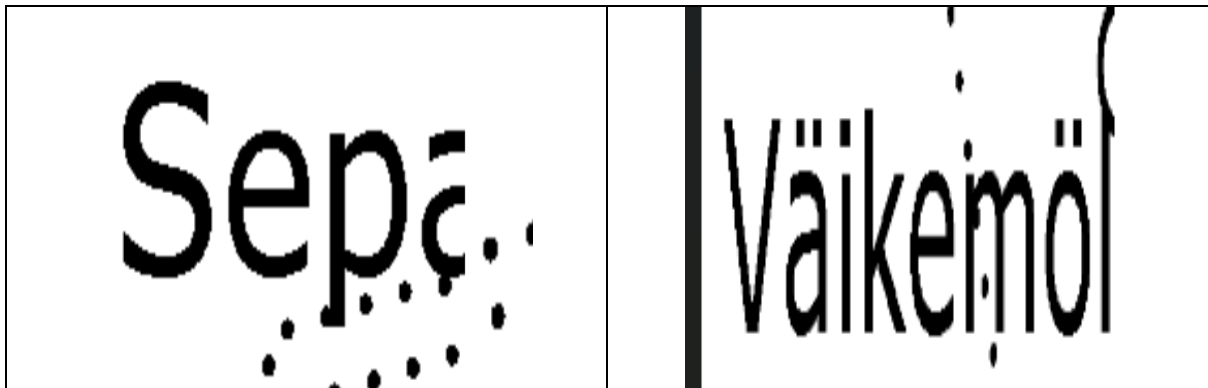
meetodi tulemus aga ka sellel juhul parem (0,99 vs. EasyOCR 0,92). Nõrgemana eristub aga Võrumaa (testala 54082) trükikaart võrreldes kombineeritud meetodit mitmemodaalse mudeliga (Gemini 3), kuna sellel EasyOCR-i saagis on madal. Mitmemodaalne mudel võibki olla parem lahendus kaartide puhul, millel on müra, kuid samas tulenevalt nõrgast saagisest on Võrumaa vanemal digitaalkaardil EasyOCR-i F1-skoor parem kui Gemini 2.5-el ja see kasvab endiselt kombineeritud meetodiga.

Tuvastades ainult keelemudeliga, seegi ei lugenud ülejäänud leppemärke nimede osaks. Kombineeritud meetodis tuleb aga mõne koha peal paremini välja tuvastuse seisukohast positiivne keelemudeli võime teha „järeldotsi“ – nime ümbritsev riskülik võib olla puuduliku ulatusega, kuid ümbritsev nimestik või keelematerjali kasutamine aitab puudujääke lahendada. Näiteid väljalõigatud nimedest on toodud joonistel 17a–d. Näiteks Võrumaa kaardilehtedel riskülikute sisse jäävad *Käpamäe* (vanem digitaalkaart, joonis 17a) ning *Servamäe* ja *Aromäe* (trükikaart). Nende läheduses asub ka korrektselt tuvastatud kõrgendiku nimi (*Käpamägi*, *Servamägi* ja *Aromägi*), eeldatavasti mille alusel talunimedele on suudetud panna õige käändelõpp. Võrdluseks Iisaku kehtival põhikaardil on tuvastatud ainult talunimi *Kivimäe*, kuna ümbruses ei ole sama nimega loodusobjekti. Iisaku trükikaardil on aga hästi eristuv ja korrektselt tuvastatud nimi *Väikevälja*, mille abil on võimalik, et kõrval olev nimi *Suurvälja* (joonis 17b) on õigesti tuvastatud. *Ööbikuoru* (joonis 17c) puhul võib eeldada keelematerjali olemasolu positiivset mõju: „u“ joonistub välja kui „a“, kuid selline vorm on eestikeelses tekstis haruldane (üksik kirjaviga) kui üldse esinev. Samasugusest täiendamisest teiseks näiteks on puuduliku algustähega nimi (kuigi samasuguse alamnurgas joonetüki suurtähti väga palju ei ole), mis on tuvastatud õigesti *Lepalaane* (joonis 17d). Samas keeleoskusest võib haruldase sõnavara puhul olla kahju – näiteks *Kriivani* on tuvastatud kui *Kuivani*, kuigi „r“ ja „i“ sulanduvad pigem n-iks kui u-ks kokku.



Joonis 17a–d. Väljalõigatud nimed a) *Käpamäe* (testala 54082, vanem digitaalkaart), b) *Suurvälja* (testala 64593, trükikaart), c) *Ööbikuoru* (testala 54082, kehtiv põhikaart) ja d) *Lepalaane* (testala 54654, trükikaart). Kaartide allikas: Maa- ja Ruumiamet (2026).

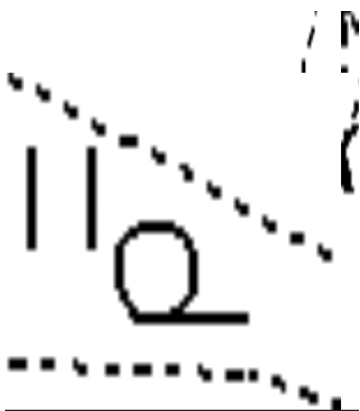
Leidub aga ka vastupidiseid näiteid nagu *Sep* (joonis 18a, testala 62471, kehtiv põhikaart). Tähestikuga tuttav inimene aimab „a“ lõpus ning kõnealune talunimi on ka tavaline (praeguste talunimede hulgas viiendal kohal 599 esinemisega – eel on järjekorras *Tamme*, *Kase*, *Mäe* ja *Saare*¹³). Joonisel 18b *Väikemõl* (testala 63644, kehtiv põhikaart) „ö“ eristub selgelt, ent on siiski tuvastatud *Väikemõl*. Ristkülikute ulatusest tulenev nime puudulik tuvastus ongi üheks probleemiks ka kombineeritud meetodi puhul, kuigi eeltoodud näidised seda kompenseerivad.



Joonis 18a–b. Puudulikult tuvastatud nimed a) *Sep* (testala 62471, kehtiv põhikaart) ja b) *Väikemõl* (testala 63644, kehtiv põhikaart). Kaartide allikas: Maa- ja Ruumiamet (2026).

Nimede õigekiri on heal tasemel ja müra kaasneb ainult üksikute nimede puhul, kui sidekriipsuga nimes kirjavahemärk on tuvastatud mõlemas osas ja ka kokkupanekul säilitatud (näiteks *Aruste--Arbo*, testala 63644, kehtiv põhikaart), mis on tingitud autori otsustest. Üksikus näites (*Uuetürnpuu*) on „rn“ tuvastatud kui „m“. Valge värviga nimed on ka tuvastatud hästi. Nime kokkusurutuse võres ei nähtud testide tulemusel avaldavat mõju tuvastuse täpsusele, kuid kindla suuruse paikapaneku asemel võib olla soovitatav säilitada kuvasuhe.

Kohati – kuid mitte alati – kui ristkülik hõlmab ainult osa nime keskelt ehk väiketähti, mudel paneb tuvastusele suure algustähe, mis välistab selle kasutamise kokkuviimise tingimusena. Näiteks testala 54082 kehtival põhikaardil on *Kõivsaare* asemel tuvastatud *Vsa*. Kui tuvastust ei suudeta teha, mudel ei paku varianti. Kui ristkülikud hõlmaksid ainult nimesid, see võiks olla problemaatiline eriti vanemate kaartide käsitsi kirjutatud nimede puhul, kuna käekiri ei tarvitse sarnaneda tuntud fontidele, kuid antud töös ei välistata sõnu vaid kaardi punktleppemärke, mis on nimedena väljalõigatud (joonisel 19 rohumaa ja harvik).



Joonis 19. Nimena väljalõigatud aga tuvastuses välistatud punktsümbolite kogum (vanem digitaalkaart, testala 54082). Kaardi allikas: Maa- ja Ruumiamet (2026).

¹³ Excelis läbiviidud arvutus. Ametlikku sagedustabelit talunimede kohta ei ole.

Kombineeritud meetodis töövoos tuli välja viiba (prompti) sõnastuse olulisus. Kui mudelil paluti ainult tagastada JSON kujul

```
[{"cell":0,"name":"..."},  
{"cell":1,"name":"..."}],
```

ehk täpsustamata, kuidas sidekriipsuga nimedega ümber käia, see liitis kokku nimesid, mis olid sidekriipsuga kuid kahe erineva võre silma sees. Põhimõtteliselt oleks loogiline nimesid niimoodi kokku panna, kuid see rikub ära nimede nummerdamise loogika, ja selle tulemusel ühendatakse nimi vale võre silmaga. Eriti kriitiline olukord on Võrumaa kaartidel, mujal vead on üksikud. Võrumaa trükikaardi peal on kolm valesti lingitud nime 20-st (15 %) ja kehtiva põhikaardi peal 10 talunime 102-st (9,8 %) on valesti lingitud. Siinkohal valesti lingituteks on loetud ka näiteks nimed *Ala* ja *Oru*, mis koos sidekriipsuga moodustavad ühe nime ja mille esimese osa jaoks on tuvastatud *Ala-Oru* kuid teise jaoks *Jöömägi*. See nimi ongi Võrumaa kehtival põhikaardil probleemi aluseks, kuna see asub võres kõrvuti silmades, aga on ühe indeksiga, mispärast mõju avaldub ka teistes sama võre nimedes.

Joonisel 20, mis on keelemudeli sisend, näitena on *Uus-Kaskemäe* (ka *Vana-Pikamäe*). *Uus-Kaskemäe* on kahes erinevas võre silmas ent algse viibaga pani keelemudel selle kokku esimese järjekorranumbriga (nummerdamine algab nullist). Siis indeksiga 1 on *Lepõllu*¹⁴ ja viga kuhjub, kui nimedes liigutakse edasi. Sellepärast oli vaja mudelile selgesõnaliselt öelda, et nimesid ei tohi kokku liita, ja nimede sulandamine tehakse eraldi sammuna.

Töös nähti vaeva, et leida sõnastus, millega mudel ei viiks omavoliliselt kokku kõrvuti asuvaid nimesid. Sõnastus saadi paremaks, kuid mitte perfektseks. Tulemustest leidub sellest asjaolust tulenevaid vigu, kuna sisendi sõnastus ja/või mudeli tegevus sisendi alusel pole perfektsed. Viga tuli hindamisel välja, kuid parandatud kujul sõnastusega viga ei kuhjunud vaid jäi puudutama üksikuid nimesid, mistõttu tuvastust uuesti läbi ei viidud. Olukord aga illustreerib hästi seda, kuidas keelemudelid on optimeeritud semantikale, mitte ruumilisele aspektile.

¹⁴ *Lepõllu* on kirjutatud õigesti, nii nagu algse viibaga tagastati, kuigi jooniselt ei ole aru saada, kust esitähht tekib.



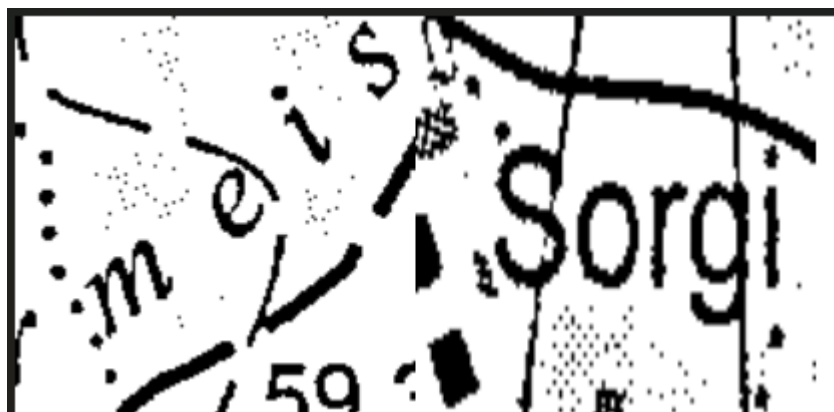
Joonis 20. Keelemudeli sisend kombineeritud meetodis ja näide nimedest, mis põhjustasid keelemudeli omaalgatuslikkust nimede kokkupanekul. Testala 63644, kehtiv põhikaart. Kaardi allikas: Maa- ja Ruumiamet (2026).

Nimede kokkupanekus on ka seletamatu loogikat. Näiteks Iisaku trükikaardil (testala 64593) *Ees-Kutsari* on pandud kokku üheks nimeks, kuid *Taga-Kutsari* puhul seda pole tehtud (vt joonis 21). Seletuseks ei saa olla see, et esiti mainitu riskülikud kattuvad üksteisega, teise aga mitte, sest näiteks *Uus-Kaskemäe* (testala 63644, kehtiv põhikaart) riskülikud ei kattu, ent nimi on siiski koos. Lisaks kui ühes võre silmas on kaks nime, on nendest töö näidetel enamasti tuvastatud ainult esimene. Erandiks on *Vahi* 47.9, mis jäi tulemustesse, kuna eemaldati ainult puhtarvulised sõned.



Joonis 21. Ees- ja Taga-Kutsari kõrvuti võre silmades. Trükikaart, testala 64593. Kaardi allikas: Maa- ja Ruumiamet (2026).

Seletamatu loogikat on ka indekseerimise tõlgenduses. Joonisel 22 toodud võres metaandmete failis on kaks indeksit: 0 ja 1. Siiski mudel on tuvastanud neli eraldi tähte igaühe oma indeksiga (*m*, *e*, *i*, *s*). *Sorgi* on tuvastatud korrektselt indeksiga 5 (number 4 on jäetud tuvastamata), aga kuna nii suuri indeksinumbreid metaandmetes ei ole, sellele on omastatud väärtus *e*.



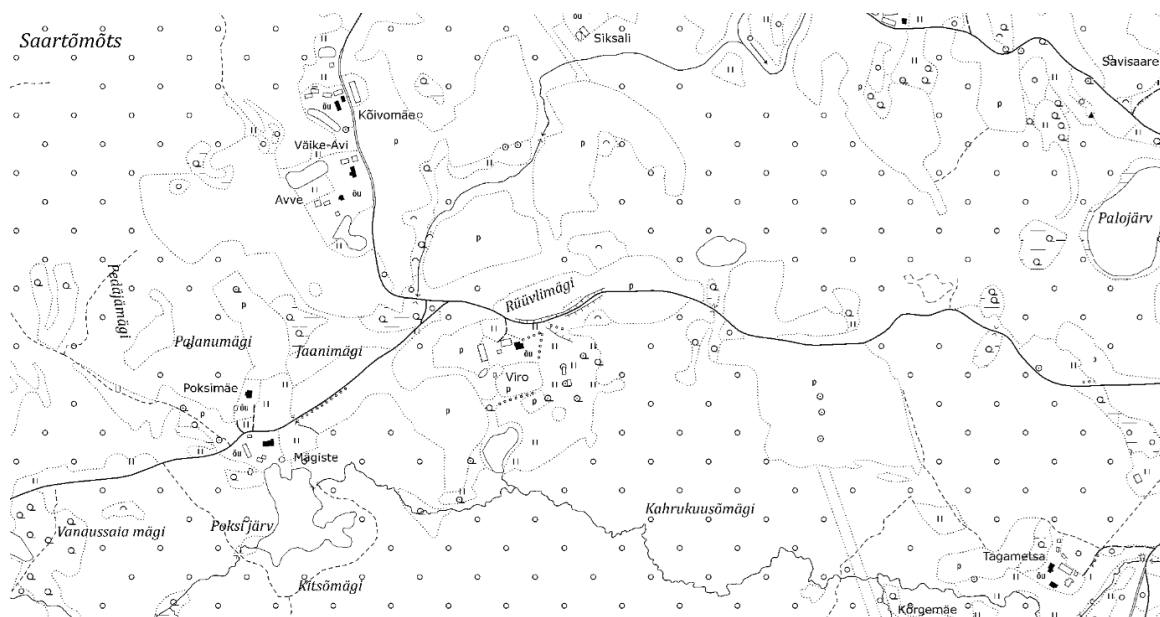
Joonis 22. Indekseerimisprobleem. Trükikaart, testala 63644. Kaardi allikas: Maa- ja Ruumiamet (2026).

Tuvastuses esinevad vead on üksikud välja arvatud eeltoodud nimede vale linkimine Võrumaa lehtedel. Selle puhul ongi soovitatav katsetada erinevaid viiba sõnastusi, aga on võimalik, et probleem täielikult lahendatud ei saa. Kombineeritud meetodi kitsaskoht on ka EasyOCR-i ristkülikute kasutamine, mis endiselt avaldab mõju kohtades, kus nimede loetavus pole ideaalne, töö näidetel eriti Võrumaal. Sellepärast kolm valesti lingitud nime Võrumaa trükikaardi peal annab suure osakaalu – nimesid on küll 50, kuid ristkülikuid ainult 20. Seega kui on kasutada mudelit, mis leiab nimed paremini üles, on kombineeritud meetod sobilik lahendus ka keerulisemate kaartide peal. Parema loetavusega kaartidel on kombineeritud meetodi tulemused samas heal tasemel.

3.1.2 Mittesobivate mudelite tulemuste hindamine

Kõik mittesobivateks liigitatud mudelid on vabavaralised: MMOCR, NewOCR, PaddleOCR ja Tesseract. Nende jaoks F1-skoori pole arvutatud.

NewOCR on oma veebiliidesega kergesti kasutatav, toetab paljusid keeli ja see võib sobida piltidele, milles tekst on sirgerealine, valge taustaga ja ilma mürata, kuid kaardi nimede lugemisel pole sellest kasu. Näiteks ei suuda see tuvastada ühtegi sõnet Võrumaa kehtiva põhikaardi mosaiikkaardilehe pealt (testala 54082). Võrdluseks testiti ka sama kaardi väljalõikega, milles nimed eristuvad paremini (joonis 23). Väljavõtte väljundist (joonis 24) sisaldab korrektseid nimesid ning ka vigaste nimede puhul teisenduskaugus pole talumatu, kuid tulemus sisaldab müra, mis ei ole reeglite alusel eemaldatav, eriti kui arvestada, et müra on iga sisendi puhul erinev. Võru keele põhjustatud raskusi pole täheldada.



Joonis 23. Lõik (ekraanitõmmis) PNG-kujul Võrumaa mosaiikkaardilehest, testala 54082. Kaardi allikas: Maa- ja Ruumiamet (2026).

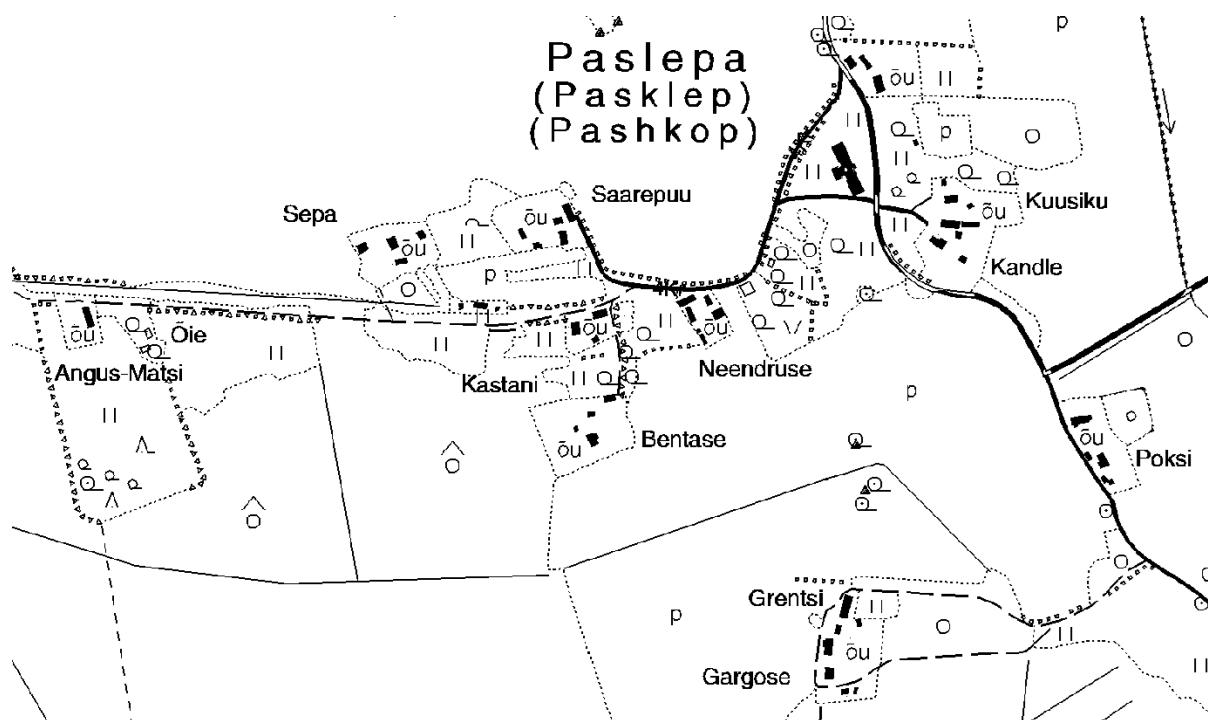
Si k .-Palanumägi a 0 a ? ; oii; c 7 NW UE -
o sete me 9: Jaanimägi 24° t > % i t ° 17 lai % ° rež N ma "2
A Sri KS * ; e ; onm GS väi mtmea, % Aan ° 11 %
ae mu, ° OOO 7 jee 40 airo 3 0 Tõi Oo 0 ° ° Tap Uta
OSMEE AS Ome A TE VIFO 3 ep ji ANA : A . KN
SO m Oole 027, Peksimäe: y : T Cc maa 0 KAKA AE pae a ; a A
TRR og Ora te Ee ESTA 00 kest a 00 7 o ° : 2 3 Oo o o Ta:
0 oo õi jt 41 N-õ Oi, ; A 0 ° o A ° ° ° 0
ra N m:) ; a si (M m i ° A — O;
° 0 SOON ma PD Om Mägist 0, 0 Oo : AA (HH os ° ° ° Ot a 0 o ° Oo 0
oe Pl > Trep i : i KR jim ve " toa veo KA n v e " ,
TOL " 0 n a mA Parma 9 7
i ° o SN O Tume 0 Tr joa 5 ° ° ° ° e) A 0 o NO ° ° m
: ra aema, * õis Ma s a 2 m A SR Ne 7
OT TA VN Mevvi Kahrkuusmägi va JUUN, /
7 t ° . Peksjärvi od M ° ° ° ° ORTO 0 e °
i : : 2 e e 1 n Su A 2 OS
Sa pp Vandussai mägi (51 JE ? J vä me
, KIK) ° o nii 77 0 ° o o o ° ° 7 3 ° peeda PO OT a
N 2; Ni 2 E s Tägametsa 2. 0 El
17 ; *s ae rT (o Pp vt 8 Ss Ar

Joonis 24. Väljavõtte NewOCR-i tuvastusest Võrumaa mosaiikkaardilehe lõigul.

Lisaks katsetati PNG-vormingus ekraanitõmmist sama ala binariseeritud trükikaardi pealt (joonis 25). Selle peal tuvastus oli saagi seisukohast olematu, ja ainus korralik tulemus on *Viro*. Võrumaa kaarte aga iseloomustab mitmekesisest maastikust tulenev rikkalik taustainfo. Sellepärast tehti katse ka testala 62471 (Noarootsi) vanema digitaalkaardi lõigu peal (joonis 26), kus on lagedamat ala ja nimed ei sulandu taustaga kokku. Tulemuste kvaliteet ühtib siiski eelmiste katsetustega (väljavõtte joonisel 27), ehk müra teeb edasise töötlemise võimatuks. Eriti küsitav on nimele (*P a S s k | e p*) järgnev pikk tähtede kett. Tulemused vastavadki esialgsetele katsetustele, kuid eriti trükikaardilt saadud väljund viitab sellele, et automatiseeritud tekstitöötlustega pole tulemusi võimalik viia kujule, mida hiljem testandmete või teiste tuvastatud nimedega kokku viia.



Joonis 25. NewOCR-i testimiseks kasutatud väljalõik (ekraanitõmmis) binariseeritud Võrumaa kaardist. Trükikaart, testala 54082. Kaardi allikas: Maa- ja Ruumiamet (2026).



Joonis 26. NewOCR-i testimiseks kasutatud väljalõik (ekraanitõmmis) binariseeritud Noarootsi kaardist. Vanem digitaalne kaart, testala 62471. Kaardi allikas: Maa- ja Ruumiamet (2026).

- KONE OS; i P

4 ? š

p | Apvirerrsegeeevod

aslepa Mpõui 11 ž

(P a Ss k | e p) Š sõtemeeageeteeieeeeeeeemenetta

(Pashkop) ö 1% pi o ž: |

ja Saarepuu £ ae

aed f + õy iKuusiku

Sepa jann ai O0* A n t mn 00

9 Või, |), on AX aso; LIIN 4 7

ao SAV ehova prerer teie i ke Pp see AN Oo põ 0 rj AR AKandle

; t Mo aaa a MEE E RA Oi C

mt AO 7 Ole mo t Ane A 27 ž °

; Ai TANA AE er Bi Neendruse x

«Angus Matsi pr Kastani *--! 5 a p mme >

z K A f 3 "3 ; Bentase 2 Nõu rd

Joonis 27. Väljavõte NewOCR-i väljundist Noarootsi vanema digitaalkaardi lõigu nimede tuvastamisel, testala 62471.

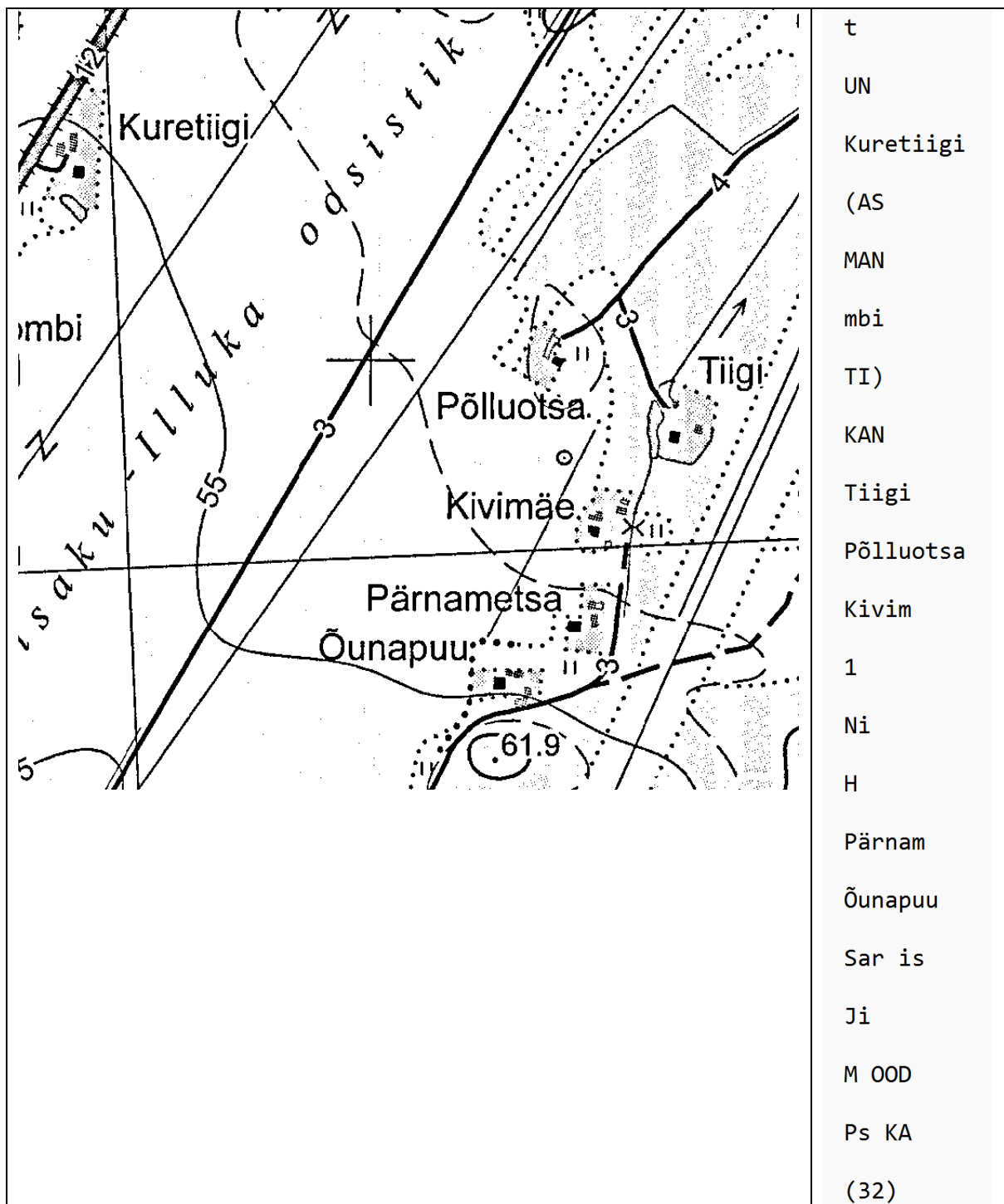
Sirgerealise, taustast hästi eristuva tekstiploki lugemiseks NewOCR aga sobib. Võrdluseks kasutati pilti varem töös viidatud Gerth Jaanimäe (2021) artikli sissejuhatuse osast (joonis 28), mille peal õnnestub eestikeelne tuvastus perfektselt, ilma ühegi veata. Isegi, kui valida tuvastuse keeleks inglise, on tekst oma väikeste vigade ja täpitähtede puudumisega inimlugeja jaoks arusaadav, kuigi näiteks ü on asendatud ti- või ii-ga (näiteks *analtitis*, mitte analüüs; *triikkima* mitte trükkima). Siiski on küsitav, milleks tulemust vaja on, kuna teksti pole enam vaja näiteks tõlkimiseks kopeerida, sest näiteks Google Lens (2026) suudab teksti pildi pealt lugeda ja selle ära tõlkida.

Ajalooliste tekstide automaatne analüüs on tänapäeval huvipakkuv ja aktuaalne suund nii keeleteadlastele, ajaloolastele, etnoloogidele, perekonnaloo uurijatele kui ka kõigile teistele, kes kasutavad oma töös digitaalsel kujul ajaloolisi tekstimaterjale. Eesti digiarhiivides on selliseid tekste praeguseks rohkem kui kunagi varem, aga enamasti on need talletatud piltidena, mida ei saa uurida automaatsete teksti-analüüsi vahendite abil. Sageli on tegemist käsikirjaliste tekstidega ja kuna käekirjad on erinevad, ei ole neid võimalik tärgtuvastuse (pildifailist automaatse trükiteksti tuvastamise) abil tekstiks teisendada. Seega peab neid tekste ümber trükkima ehk masinloetavaks muutma inimene.

Rahvahanke korras on sel moel Rahvusarhiivi andmebaasi sisestatud rohkem kui 55 000 käsikirjalist eestikeelset vallakohtu protokoll, mis sisaldavad rohkem kui 2,5 mln tekstisõna¹ ning tärgtuvastuse abil on digiteeritud vanu ajalehetekste. Kuigi tärgtuvastusega on tekstide masinloetavaks muutmine odavam ja kiirem, on probleemiks tekstituvastuse käigus tekkinud vigade suur arv. Vanade tekstide automaatseks lingvistiliseks analüüsimiseks tihti ei piisa nende masinloetavaks muutmisest.

Joonis 28. Pilt G. Jaanimäe (2021: 47) artiklist, sisend NewOCR-ile. Raam autori lisatud.

Tesseracti tulemuste suurim mure on, nagu NewOCR-i puhulgi, müra, mis teeb tulemuse edasise töötlemise raskeks, kuigi seda on NewOCR-ist vähem. Ta lubab pisut rohkem paindlikkust ja saagise tõstmiseks palutigi sellel otsida mitte ainult potentsiaalseid tekstikaste vaid lihtsalt lugeda kõik, mida tekstiks hinnati (--psm 11, vt Tessdoc, kuupäev puudub). Müra pärast oluline nõue ongi, et tagastatakse võimalikult palju teksti, mis annab iga tuvastustulemuse omaette reale. Tesseractiga katsetati nii tervikpilte kui ka ruutudeks jagatud pilte. Tervikpiltide tulemused erinevad ruutudeks jagatud piltidest ainult arvuliselt, mitte kvaliteedi poolest. Testitud trükikaardi tulemused on Tesseracti puhul kehvema kvaliteediga nii täpsuse kui ka saagise poolest kui kehtiva põhikaardi. Kuna tulemused sarnanevad NewOCR-i omadega ja sarnaselt ka Tesseracti saab eeldada sobivat selgetele, mürata piltidele, siinkohal on toodud ainult üks näidis, mis on inimsilma jaoks selge nimestikuga (joonis 29a). Selle tuvastuse täpsus on teistest näidetest suurem, ent ka selle tulemustes on väärpositiivseid tulemusi (joonis 29b). Siiski tulemusi pole püütud kokku viia testandmetega, mis oleks olnud võimalus väärpositiivseid välja filtreerida, mis aga oleks nõudnud eraldi lähenemist, kuna tulemused on tekstina.



Joonis 29a–b. a) Sisend Tesseractile, ruut Iisaku trükikaardist, testala 64593. Kaardi allikas: Maa- ja Ruumiamet (2026). b) Tesseracti tulemus Iisaku trükikaardi ruudult, testala 64593.

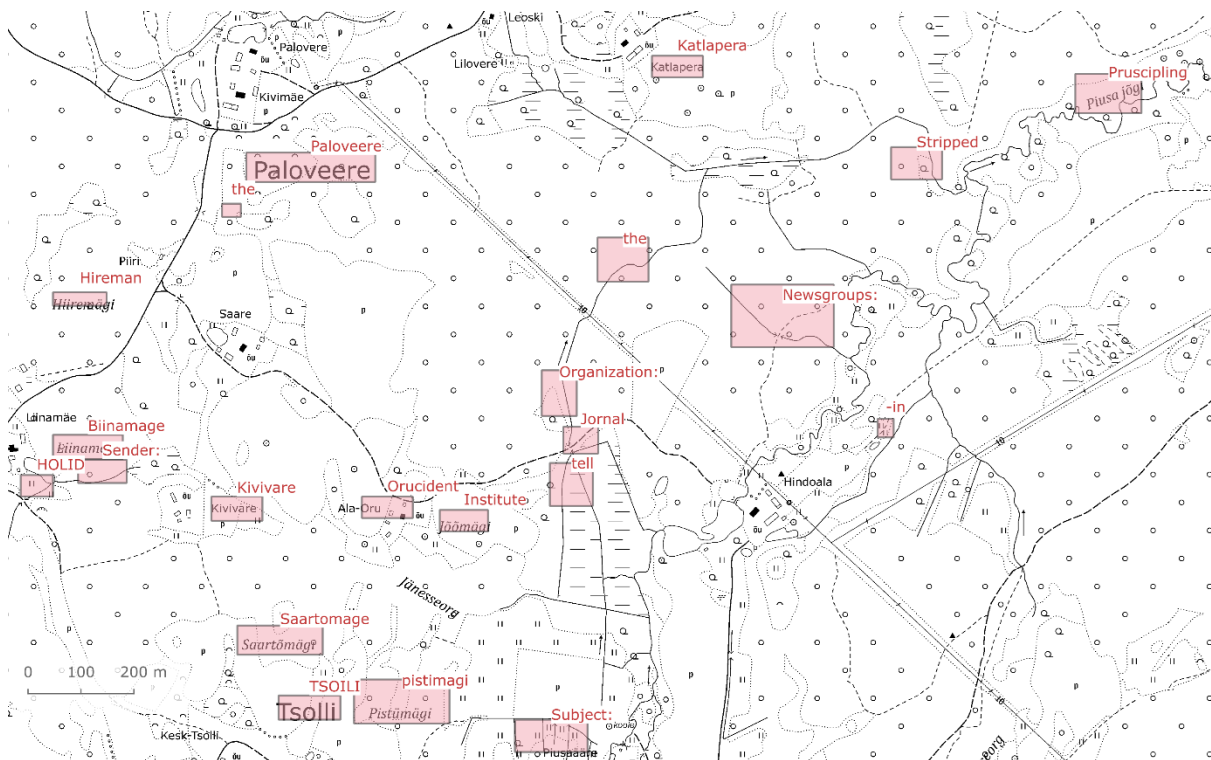
Alates neljandast versioonist saab Tesseractis kasutada LSTM-i (*Long Short-Term Memory*) ehk ühte tüüpi närvivõrku (Tessdoc, kuupäev puudub), kuid tulemused on näidistel samad kui seda eraldi kutsumata. Selle tuvastustulemused ühtivadki kvaliteedi poolest endiselt Christian Haacki (2018) magistritöö tulemustega, ehk ka käesoleva töö andmete näitel saab temaga nõustuda, et Tesseract professionaalseks kasutamiseks ei sobi. Haack (2018) kasutas nimede kokkuviimisel kohanime loendit, ent ka siis tuvastus jäi nõrgaks, ja eriti talunimede puhul see lähenemine ei tööta, kuna ei ole olemas võrdlusandmestikku. Kaartide nõrkadele

tulemustele vaatamata saab Tesseract sarnaselt NewOCR-ile Jaanimäe (2021) artiklist võetud näidispildiga hästi hakkama (vt lk 58).

MMOCR on paljusid mudeleid sisaldav tööriistakast, mispärast tulemused võivad suurelt sõltuda mudelite valikust. Antud töös on valitud leidmiseks DBNet++ ja tuvastuseks SAR (vt täpsemalt ptk 2.4.2 Testide läbiviimine). Leidmiseks kasutatud mudeli küsitavaim koht on, et see tagastab riskülikuid, mille sisu ei vasta ladina tähestikule. Tuvastusmudel omapoolt töötab tugevalt inglise keele põhjal.

Joonisel 30 on väljalõik leidmise ja tuvastuse tulemustest kehtiva põhikaardi Võrumaa mosaiikkaardilehel, mis valiti sellepärast, et sellel esineb palju mittehorisontaalseid nimesid. Sellel on näha vahelduvust tekstide leidmise suhtes: mõnel juhul on leitud õigesti, mõnel juhul inimsilmaga risküliku sees teksti pole näha ning paljudel juhtudel nime pole üldsegi leitud, mis teeb saagise madalaks. Mõni leidmata nimi on ka selgel valgel taustal (nt *Saare, Hindoala*). Antud juhul pilti pole jagatud ruutudeks ja sellepärast riskülikud väga palju üksteisega ei kattu. Siiski terve kaardilehe paremas alamnurgas on ala, millelt pole leitud mitte ühtegi nime.

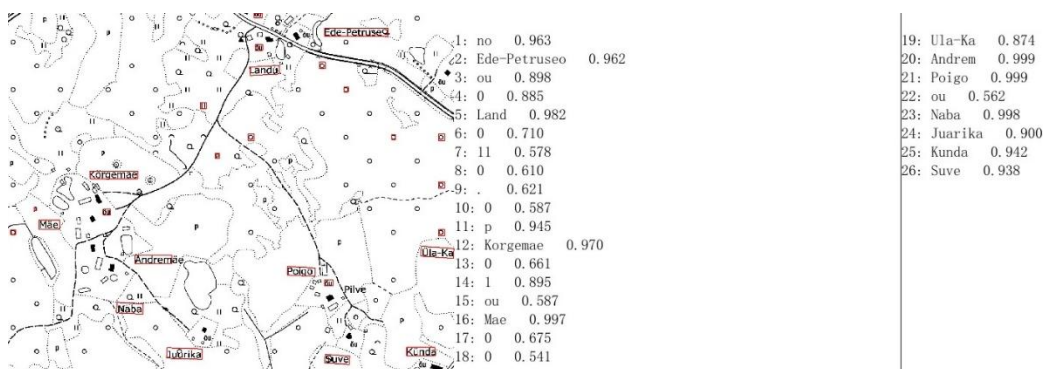
Tuvastamisega problemaatiline on see, et alati pole teada, mis keeli mudelid toetavad, ja tulemustest on näha, et valitud mudel töötab inglise keele põhjal. Täpikätesid üldse ei tuvastata. Lisaks on suhe riskülikus nähtava ja olemasoleva sõnavara vahel hägus. Inglisekeelsete tulemuste hulgas on näiteks *Partiture* (vaste *Palojärvi*), *Institution* (vaste *Jõõmägi*) ja *industrial* (vaste *Rüüvlimägi*). Kaardi joonleppemärke on tõlgendatud inglisekeelsete sõnadena, näiteks teejoon ja punktiirjoon (mõlemad *the*), jõe joon- ja metsa punktleppemärk koos (*Sender:*) ning kraavi- ja punktiirjoone leppemärgid (*Organization:* – enesekindlusega 0,889). Tulemused on samasugused iga testitud kaardilehe peal. Mõne nime puhul jääb mulje, et mudel on sunniviisiliselt tuvastanud midagi, mis on inglisekeelne või sellele sarnane sõna: näiteks nimest *Ala-Oru* on leitud järgmine osa ja selle alusel tuvastatud sõne *Orucident* ja nime *Hiirimägi* alusel on tuvastatud *Hireman*. Tulemused reetvad ära, et neid ei saa automaattöötusega ära eestistada. Suuremate ressursside puhul saaks mudelit eestikeelsete nimedega treenida ning on võimalik, et installimisega on kaasnenud ka mudel, mis oleks vähemal määral tingitud inglise keelest. Tuvastustulemuste alusel võib arvata, et inglisekeelse andmestikuga tasuks katsetada, ent eestikeelsete kaardiandmete jaoks valitud mudelid ei anna rahuldavat tulemust.



Joonis 30. Väljalõik MMOCR-i tulemustest Võrumaal. Kehtiv põhikaart, testala 54082. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

PaddleOCR tuvastab ka inglise keele alusel, sest eesti keelt ei toetata, ent ei suru inglisekeelseid sõnu peale nagu MMOCR-i mudel, kuigi täpitahtesid kohati ei tuvastata. Selle suurim mure on aga käesoleva töö testide tulemustel koordinaatide viga.

PaddleOCR VL-mudeli sisaldav versioon pakub veebis demot¹⁵, mille tulemus andis alust arvata, et edasitoimetamine on põhjendatud. Joonisel 31 on toodud demo tuvastustulemus kehtiva põhikaardi väljalõigult testalalt 54082. PaddleOCR-i skript on keerulisem kui EasyOCR-i oma, kuid tulemustes on vähem müra.



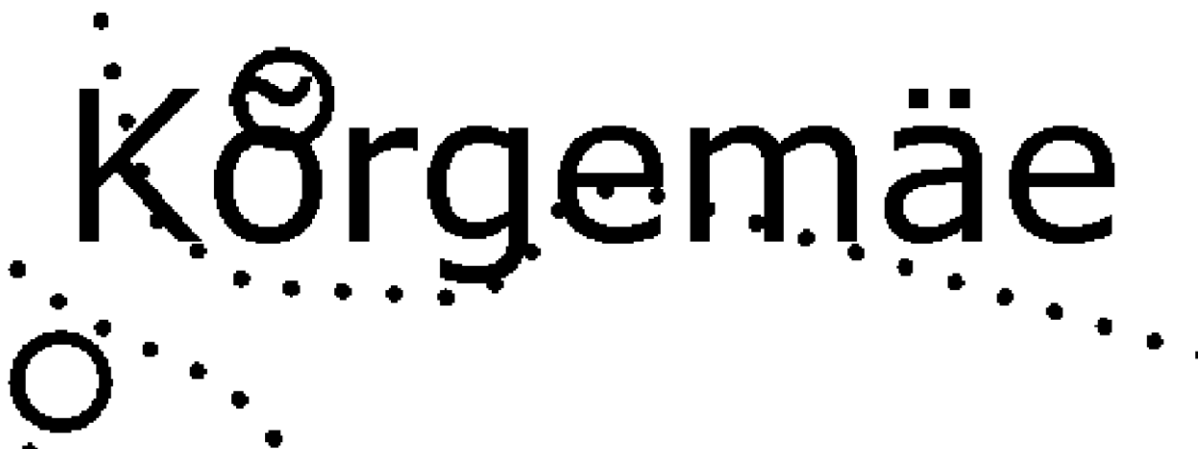
Joonis 31. Paddle-demo tuvastuse väljund.

Demo ei reasta tuvastusi enesekindluse alusel vaid järjekorraks on tuvastuse järjekord. Jooniselt 31 on näha, et tekste ümbritsevad riskülikud on enamasti joonistatud õigesti ning tuvastuse täpsus on adekvaatsel tasemel ka seda numbriliselt mõõtmata, kuigi täpitahti pole

¹⁵ <https://huggingface.co/spaces/PaddlePaddle/PaddleOCR-VL> Online Demo Tulemus on saadud 7.1.2026. Hilisemal katsetamisel 1.4.26 kasutajaliides on muutunud: riskülikuid ei joonistata ja tulemused viitavad veale.

tagastatud. Üks nimi (*Pilve*) on jäänud tuvastamata, mis on nõrkus, kuid võib olla ka üksiku sisendpildi puhul esinev viga ja oma iseloomult pigem anomaalia.

Pythonis testiti edasi vanema ehk mitte keelemudelit sisaldava versiooniga sama testala mosaiikkaardilehega, mis jagati tuvastuse käigus väiksemateks ruutudeks. Kuigi uuema versiooni demo andis adekvaatse tulemuse, ilma keelemudelita versioon jääb hätta. Tuvastuse tulemuseks oli 1959 sõnet. Leidmisel probleem on see, et paljud tekste ümbritsevad riskülikud ei katu tekstidega korralikult või üldse, ent kattuvad üksteisega. Seega ei saa neid kasutada sisendiks paremini tuvastavale programmile. Kui ühe nime kohta käib mitu riskülikut, oleks lahenduseks sulandada need kokku rakendades reegleid (vt Schlegel, 2021), aga sellega kaasneks müra ning selge on, et nimed tervelt riskülikute sisse ei jää. Suurema pildi puhul tuli välja, et inglise keelele vaatamata täpitähtedest ä ja ü olid enamasti tuvastatud õigesti (*Hirdemäe*, *Künda*) ent kohati ka mitte (*Magiste*, *Saagrimae*), mis ei paista jälgivat reeglit. Õ asemel oli enamasti tuvastatud ö (*Köivsaare*), kohati ka o (*Vosu*) ning erandina tulenevalt taustmürast ka ø (joonis 32).



Joonis 32. Taustmüraga *Kõrgemäe*, PaddleOCR tuvastatud *K8rgemäe*. Kaardi allikas: Maa- ja Ruumiamet (2026).

Kriitiline on ka joonisel 33 toodud olukord, et tuvastatud nime punkt ja ümbritsev riskülik ei oma seost testandmestiku riskülikutega. Näiteks *Potsepa* punkt asetub testandmestiku risküliku sisse, kuid tuvastatud *Meelaku*, *Reik* ja *!ssn[* jäävad riskülikutest välja – lisaks erinevates suundades ja kaugele sellest objektist, millele nad viitavad – ja sellepärast ei saa neid ruumipäringuga kokku viia vaid see tuleb käsitsi teha, mis suurema andmemahu puhul pole rakendatav lähenemine. Mõningaid tulemusi oli ka kaardilehest väljapool ida- ja lõunaservades, mis viitab koordinaatide probleemile, kuid ka see ei näidanud süstemaatilist viga. Lisaks joonisel on näha puhtal valgel alal olevaid riskülikuid, milles kohati on tuvastatud tekst (näiteks *no*), kohati mitte. Teine näidis on toodud joonisel 34, kus *Nostalgia* on tuvastatud kahel erineval juhul, kuid nime *Vilmani* pole üldse tuvastatud ning *JODSS* ja *Fohunins* tunduvad hallutsineerimise tulemus, kuigi OCR-mudelitega see ei peaks olema suur probleem. Testimise jaoks viidi tuvastatud sõned siiski käsitsi nimedega kokku. Arvestades olukorda, millest joonised 33 ja 34 on näidisteks, ei saa välistada vigu, kuna mõni tuvastatud nimi võis olla ka kaugemal testandmestiku riskülikutest ja/või nii moondunud kujul, et autor ei teadnud neid kokku viia. Näiteks nimega *Vilmani* (joonis 34) on kokku viidud ruumilise vahemaa pärast ja teise kandidaadi puudumisel *Nostalgia*. *Vilmani* juures asuva sõne tuvastusele antakse isegi pisut suurem enesekindlus (0,992 vs. 0,979).

PaddleOCR tuvastab paljusid punktleppemärke sõnedena ja tulemustes on palju tühjasid ja lühikesi sõnesid.) Need on samas suurusjärgus EasyOCR-i tulemustega, mis olid samal lehel 50,98 % ja 73,52 %, kuid EasyOCR säilitab koordinaadid ja PaddleOCR-i arvud on käsitöö tulemus. Võrumaa kaardi keerulisus ei oma saagise seisukohast tähtsust, sest näiteks Noarootsi kehtival põhikaardil (testala 62471) ruumipäringuga saadi vaste 28-le nimele 83-st ehk 33,73 %-le. Ka siinkohal on välja jäetud tulemused pikkusega alla kolm. Iisaku vanemal digitaalkaardil, mille peal muidu ainsana saavutati kombineeritud meetodiga F1-skoor 1,0 ning ka ainuüksi EasyOCR-iga puhastatud tulemustes (teisenduskaugus 0 või 1) skoor on 0,98, olukord on testitute kõige kehvem: tuvastustulemustest pikkusega vähemalt kolm saadi ruumipäringuga testandmetele üks vaste 44-st ehk 2,27 % ning ka see on täiesti vale koha peal.

Täpsust saaks edasise sõnetöötusega tõsta eemaldades tulemustest kirjavahemärgid nagu *..Kesk:-Tsolti-* (p.o *Kesk-Tsolli*), kuid seda pole antud töös tehtud tulenevalt tuvastuste valedest asukohtadest. Vahemaad nime ja tuvastustulemuse asukoha vahel on kohati suured, mispärast pole kindel, kas tulemusi saaks keelelise sarnasuse ja ruumipäringuga kokku viia praeguste talunimedega, ja süstemaatilise nihke puudumisel see tõenäoliselt sisaldaks palju vigu. Töös saadud tuvastustulemuste alusel katsetamine poleks vaeva väärt ja kuna käsitöö on suures osas andmetöötusel, pole PaddleOCR põhikaartide tuvastusel arvestatav variant.

3.2 Nimekihistute kokkuviimise hindamine

Saamaks teavet võimalustest erinevate aegade nimesid kokku viia, kasutati Levenshteini teisenduskaugust erinevate lävendite ja kahe erineva lähenemisega. Aluseks olid kombineeritud meetodiga saadud tulemused, kuna nende puhul nii täpsus kui ka saagis olid paremal tasemel kui üksikul mudelil. Kõigepealt viidi iga nimekiht kokku praeguste talunimedega ja siis viidi iga ajalise kihi tulemused kokku vanema kihiga. Igale punktile loodi 250 m puhver, mille seest otsiti viis lähimat vastet. Nendest valiti välja parimaks kandidaadiks see, mille teisenduskaugus võrdlusnimega oli kõige väiksem, ja kui selliseid oli mitu, valiti see, mis oli ruumiliselt kõige lähem. Hindamises modelleeriti tegelikku olukorda viies kokku tuvastustulemusi, mitte testandmeid. Arvulised tulemused on toodud tabelis 4 ja joonisel 35.

Praeguseid talunimesid on andmetes 489 tk. Võrreldes kehtiva põhikaardi testandmetega kogus on 82 nime võrra suurem. Neid nimesid aga ei eemaldatud andmetest, sest need võivad esineda vanematel kaartidel kas talunimena või muu asustusüksuse nimena. Siiski tuleb selle lõhega hindamisel arvestada ehk arvulised tulemused annavad esialgse hinnangu. See, kas talunime linkimine asustusüksuse nimega on õige või vale, ei ole üheselt mõistetav, kuid antud töös on see hinnatud valeks. Korrektselt kokkuviidud nimede hindamist on praegu tehtud osaliselt ning täpsete mõõdikute nagu saagis ja täpsus arvutamiseks tuleks teha kindlaks kõik kokkuviidavad ning ka ilma vasteteta talunimed, aga see jääb töö mahtu arvestades käesolevast tööst välja.

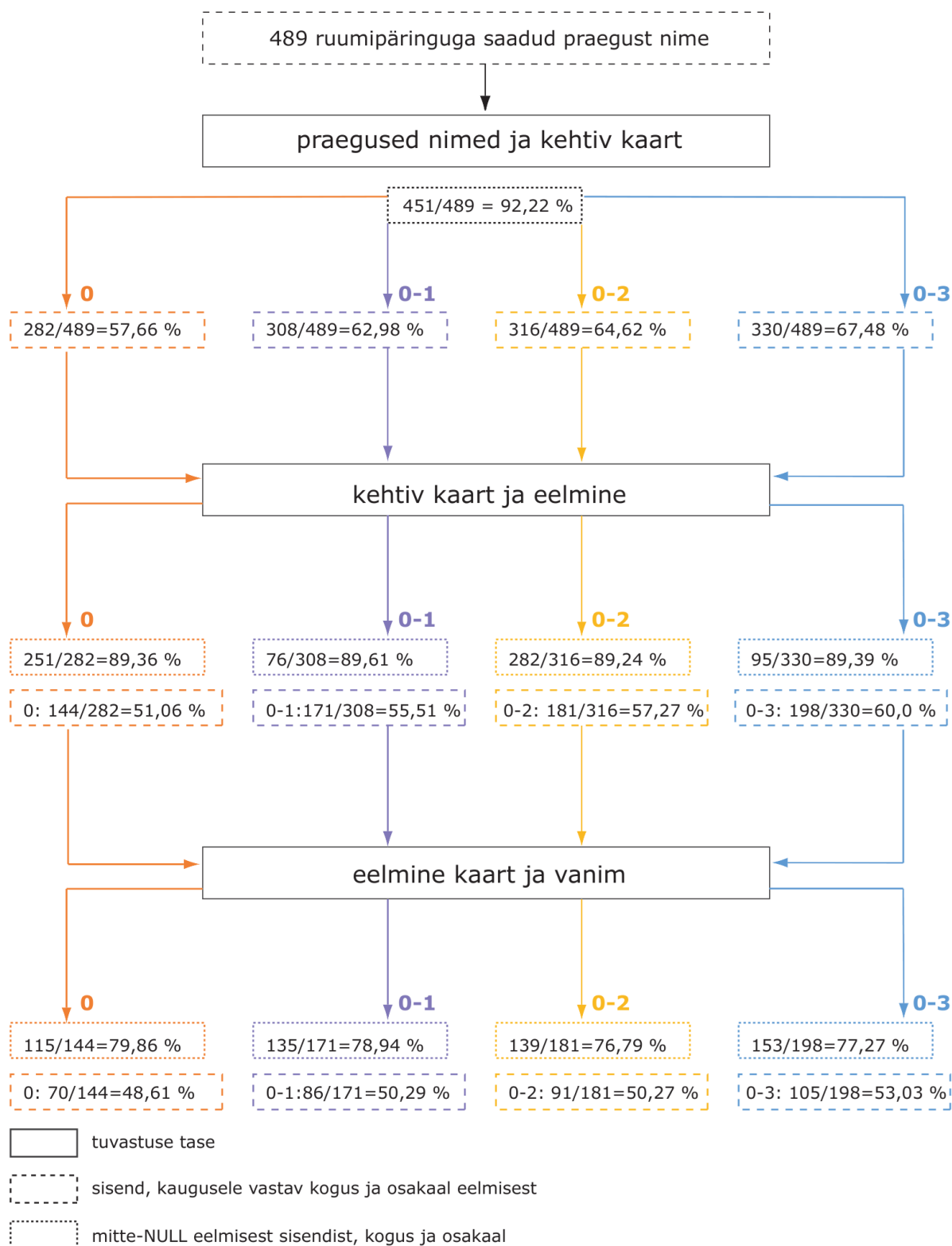
Tabel 4. Ajaliste nimekihistute kokkuviimise arvulised näidud, referentsiks praegused nimed.

Kaugus →	0	%	0–1	%	0–2	%	0–3	%	NULL	%
Aeg ↓										
kehtiv	282	57,66	308	62,98	316	64,62	330	67,48	38	7,77
keskmine	211	43,15	221	45,19	233	47,65	248	50,72	77	15,75
vanim	128	26,18	136	27,81	140	28,63	158	32,31	133	27,20

Iga kihi kokkuviimist praeguste nimedega eraldi tehti mõistmaks, mis nimed ja kui suur osa nimedest on püsinud sama. Mainitud lõhe ning tuvastusmudeli nõrkused tulevad välja selles, et teisenduskaugusega 0 kehtiva põhikaardi andmetestki ainult 57,67 %-le leiti vaste. Mitmemodaalsete mudelite hindamise puhul on toodud välja nende hea täpsus, ja kihistute kokkuviimises peegeldubki EasyOCR-tuvastusmudeli puudulik sõne leidmine ehk nime algusest ja/või lõpust puudub täht(i). Seda näitab vastete osakaalu kasv tõstes teisenduskauguse lävendit. Vaadates teisenduskaugusega 3 kokku pandud kehtiva kaardi nimesid (14 tk), kaks nendest on valesti lingitud, sest kaardil kirjas olevat nime pole tuvastatud, kuid nelja nendest kaardil ei esinegi. Ülejäänud on õigesti. Nendest 121 nimest, mille teisenduskaugus praeguse nimega on suurem kui 3, 16 nime on lingitud õigesti, aga tuvastus on puudulik. Suurima kaugusega õigesti lingitud on *gam*, mis on osatuvastus nimest *Tagametsa*, kaugusega 6.

Keskmisel kihil kaugusega 3 ainult kaks nime 14-st on õigesti lingitud. Piiripealsed juhud on praegused nimed *Saagrimäe* ja *Vabrikusaare*, millest esimene on lingitud külanime ja teine saare nimega. Väiksemate kaugustega nimedest leidub kaks valesti lingitud nime. Lisaks teist tüüpi objektiga linkimisele ka õigesti lingitud nimede puhul põhjused on mitmekesised: esineb kirjavigu tuvastuses (*Uus-Urge* asemel *Jus-Urge*), puudulikku tuvastust (*Saviküünka* asemel *viküünka*) ning muutunud kirjalpilt (praeguse *Võro-Ando* asemel oli *Verro-Ando*).

Vanemate kihtide puhul saab täheldada mõnede nimede säilimist sama kujul. Vanima kihi puhul kuni lävendini 0–2 on lingid õigesti ja selles klassis suuremate kauguste puhul põhjuseks on peamiselt muutused kirjalpildis (praegune *Sobliku* oli *Soobliku* ja praegune *Metsa* oli *Mõtsa* – järgmise näite üldkeelestumine on Võrumaa testalal olnud laiemgi tendents), puudulik tuvastus on ainult *urmõmäe* (p.o *Nurmõmäe*). Siiski näiteks vanima kihi 3-kaugusega nimedest (18 tk) ainult kaks on korralikult lingitud: *Nygorsi* (praeguselt *Nygårds*) ning *Vahtra* (praeguselt *Vahtramäe*). Juba need näited illustreerivad oma mitmekesisuses kokkuviimise raskust, kuna aja jooksul tekkinud vahed on täiesti erinevat tüüpi. Samane *Tagamõtsa* on ka vanimal kihil suurima kaugusega 8 (tuvastatud *a*, vt joonis 11a) aga sama kaugusega on ka *Saareke*, mis nüüdseks on *Väike-Saare*. Ühe väiksema kaugusega õigesti lingitud on ka *Mallikamäe*, mis vanimal kaardil on *Mäe-Mallika*, ning praegune *Aruste-Jaani* ja vanima kaardi *Jaani*. (*Mallikamäe* näites ei aita ka üldistatud teisenduskaugus, sest see ei luba tõsta eesliidest järelliideseks või vastupidi, vaid sarnasust tuleks otsida ühe või teise alusel.) Samas pole ka üheselt mõistetav, millal link on õige: näiteks vanima kaardi *Kruusiaugu* on lingitud kahe praeguse nimega *Suure-Kruusiaugu* ja *Väike-Kruusiaugu*. Eeltoodu alusel nime, mis on hiljem saanud hargtäiendi, on käsitletud samana, ent probleem muutub teiseks, kui võimalikke vasteid on mitu. Samuti kahese kaugusega piiripealne juhtum on praegune *Pargi*, mis vanimal kaardil on roheala (*park*).



Joonis 35. Ajaliste nimekihistute kokkuviimise arvulised näidud, referentsiks eelmine kiht.

Joonis 35 illustreerib nimede säilimist võrreldes eelmise ajalise kihiga. Seega ülim tase sisaldab sama informatsiooni, mis on toodud tabelis 4. Ka selle lähenemise eesmärk oli saada kätte need nimed, mille kuju ei ole muutunud, kuid siin avaldab tuvastustulemus suuremat mõju kui praeguste nimedega võrreldes, milles ainult võrdluse teine pool on tuvastuse väljund.

Eelmise kihiga võrreldes 0-kauguse protsendid on suuremad ja kogused väiksemad kui võrreldes iga kihti eraldi praeguste nimedega, kuna osakaalud arvutatakse väiksemast tervikust ja kogustes on näha OCR-i puudujääke. Näiteks praegused nimed *Ando* ja *Andu* on olemas praegustes nimedes, kehtival põhikaardil ja eelmise ajalise kihi kaardil (antud juhul testala 54082 vanem digitaalkaart). Seega need on olemas ka võrreldes eelmise kihi nimistut praeguste nimedega. EasyOCR aga on tuvastanud need kehtivalt kaardilt puudulikult – *Andu* kujul *ndo* ja *Ando* mitte üldse – ja sellepärast nende teisenduskaugused on vastavalt 2 ja 7, sest järgmisega on lingitud puhvri seest täiesti erinev nimi. Sellepärast pole need ka sisendi nimede hulgas. Teisisõnu joonisel toodud arvud peegeldavad tuvastuse tulemuslikkust.

Et trükikaardil võiks oodata nimede tiheduse olevat hõredam mõõtkavast tuleneva generaliseerimise astme tõttu, vähemalt nende andmete alusel eelmise kihiga kokkuviiimisel ei kajastu. Ilma vasteta tulemusi on küll näiteks Noarootsi trükikaardil (testala 62471) 10, mis on testandmetest 21,73 %, kuid teiseks kõrgem osakaal on juhul, milles eelmine kiht on digitaalkaart (14,03 %, Kiili, testala 63644), ja Iisaku trükikaardil (testala 64593) ilma vasteta tulemusi ei esine.

Suuremate kauguse lävenditega tuvastuse tulemus avaldab endiselt mõju, ent esineb ka teisi põhjuseid. Näiteks kui võrrelda lävendi 0–3 vasteid kaugusega 0, on neid praeguste nimedega võrreldes 211 ja eelmise kihiga võrreldes 147. *Ees-Kutsari* on olemas nii kehtival kaardil kui eelmisel (trükikaart, testala 64593), kuid kuna seda pole kehtivalt kaardilt tuvastatud, ei saa see esineda ka eelmisel kihil. *Haavistu* (vanem digitaalkaart, testala 54082) jääb sellest hulgast välja ainuüksi ühese kauguse alusel ja *Jagas* ei esine kehtival kaardil ent esineb eelmisel (trükikaart, testala 62471) ja on selle pealt ka õigesti tuvastatud.

Puhtarvuline vaatlus ei saa aga välja tuua tegeliku nimistu mitmekesisust ja -kihilisust. Eelpool on toodud näiteid sellest, kuidas ka see piir on hägus, kas nimi on õigesti lingitud või mitte. Lisaks esineb nimeahelaid ja nimesiiret. Ahelast näitena olgu Võrumaa (testala 54082) vanima kaardi *Oro*, mis uuemas kihistus on saanud *Saare* rööpnimeks. *Saare* on kehtival põhikaardil kinnitunud põhinimeks, aga kagu poole asuva talu nimeks on saanud *Ala-Oru*. Sama koha peal on hoonestust ka vanematel kaartidel, ent nimena on toodud ainult kõrval asuva künga nimi (*Jõõmägi*). Samal alal nimesiirdest näitena on *Halla*, mis esineb mõlema vanema kaardi peal. Täna on sama koha peal *Metsavahi*. Kuigi läheduses on *Halla mets*, nime tähendus tuleb välja alles vaadates vanemat kaarti, sest eestiaegsel topograafilisel kaardil asus selles kohas *Halla metsavaht*. See näitlikustab ka seda, kuidas uuemate kaartide analüüs võib puudulikuks jääda ja teisest küljest seda, kuidas vanema kihistu kaasamine annab juurde väärtuslikku uuematel kaartidel kajastumata teavet.

Kokkuvõttes eeldaks nimede kokkuviiimine tõhusamat tuvastust nii saagise kui ka täpsuse poolest. Üldistatud teisenduskaugus lahendaks harg- ja tagatäiendite teema ära (ja aitaks ka üldisemalt ees-, järel- ja siseliidetega), kuigi suurema andmemahu puhul eeldab see põhjalikku tööd, et leida kõik võimalikud variandid, sest nagu eelpool toodud näited illustreerivad, nende puhul muutused on mitmekesised. Kuna selle abil saab leida ka affikse, võiks üldistatud teisenduskaugus aga aidata sarjanimede ühendamisel, kui need sisaldavad sama osa. Näiteks vanimal Tartu ümbruse (testala 54654) kaardi peal on talunimi *Mätta*, mis on ka järgmise aasta trükikaardil, kuid selle läheduses asuvad hooned on saanud kaardile nimed *Haldjamätta*, *Haldja* ja *Väikehaldja*. Kehtivale kaardile on jäänud eeltoodud nimedest ainult *Haldja*, kuid *Väikehaldja* on endiselt talu nimi, mis on leitav praegustest talunimedest, kuigi seda kaardil pole kujutatud. Nimede muutumise mitmekesisus eeldakski veel mingite teistsuguste võtete

rakendamist, sest ainuüksi teisenduskauguse lävendi rakendamine tähendab paratamatult seda, et kas õigeid vasteid jääb välja või et õigetega kaasneb ka väärpositiivseid, kuigi üldistatud teisenduskaugusega vähem kui tavalise (vt Orasmaa et al., 2010). On ka võimalik, et talunimedele kihistute kokkuviimise tööahel ei saa olla täisautomaatne, vaid et selle abil saadakse kandidaadid, mis tuleb lõpuks inimtööga üle vaadata. Sellel juhul tulebki kaalutleda, kas kulukam on otsida lisateisendusi ja need töövoosse lisada või sõeluda tulemusi jätmaks õiged variandid.

3.3 Metoodika hinnang ja kitsaskohad

Tulemustes (3.1 Testide tulemuste hindamine) on toodud välja probleeme, mis on spetsiifilised ühele või teisele tuvastusmodelile ja siin käsitletakse teemat üldisemal tasemel. Lõplikku tuvastusse valitud mudelite tööd saab pidada heaks, kuigi eranditena joonistuvad välja reljeefi poolest rikka Võrumaa nõrgemad tulemused. Adekvaatne võrdlus on inimtöö. Sellega võiks saavutada paremaid tulemusi saagise poolest – täpsus on mitmemodaalsetel mudelitel heal tasemel, kuid ka inimeste puhul ei saa sajaprotsentilisele saagisele loota – ja eriti erinevate ajaliste kihistute kokkuviimisel, kuna kõik nimemuutused ei allu selgetele teisenduskauguse abil väljendavatele reeglitele. Inimtöö on igal juhul kordades aeglasem ja seega ka kallim.

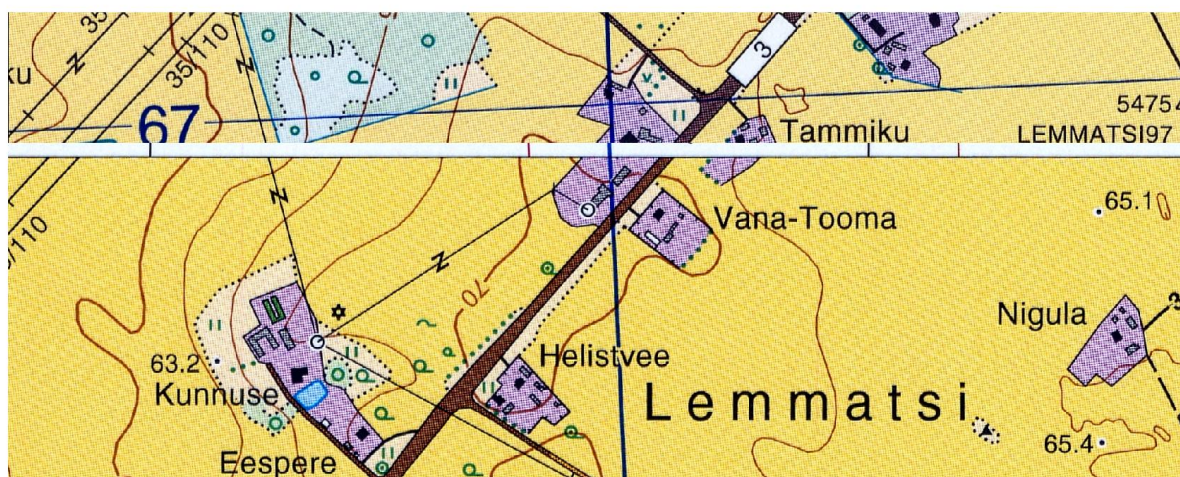
Tuvastuse tulemused sõltuvad valitud mudelist ja ilmne on, et tehnoloogia arenedes hoogsalt ja ühe magistritöö raames ei saa läbi katsetada kõiki tuvastusmudeleid. Seega teiste mudelite valikul oleks võinud saavutada teistsuguse tulemuse. Leidub aga mõningaid üldisemaid aspekte, mis kas võivad avaldada mõju uute andmetega tuvastuse läbiviimisel või mille abil tulemuslikkust oleks saanud parandada. Osaliselt need kaks aspekti võivad ka kattuda.

EasyOCR või sellele toetuv kombineeritud metoodika ei tööta suure sõrendusega nimede puhul, sest tuvastusmudelitel on raske aru saada sellest, et tegemist pole üksikute tähtedega. Taolisi nimi on näiteks Noarootsis (testala 62471) poolsaare nimi (vt joonis 13). Selliseks otstarbeks tuleks uurida näiteks Olson et al. (2024) vähima ulatuse puu lahendust, aga piirdudes talunimedega võib selle tähelepaneku kõrvale jätta, sest talunimesid sõrendusega kirja ei panda.

Vastavalt kaartidel tehtud pistelistele mõõtmistele tekitati kaardilehtedele 200 m puhver lähtudes testalade lehtede nimede paigutusest. Tervel Eesti alal ja kõikidel kaartidel mõõtmisi läbi viia ei saa, ja pole välistatud, et laiem puhver oleks põhjendatud. Sarnast metoodikat kasutades on soovitus optimaalse puhvri laius üle kontrollida. Kui teha tuvastust ühe lehe alal, peamine on, et puhvri sisse jääksid kõik nimed, mille viidatud objekt asub kaardilehe sees. Teisest küljest tihtipeale mingi nimi katkeb servas ära sõltumata puhvri laiusest.

Kuna binariseerimise protsessis valge värviga nimed kaovad, oleks testandmestik tulnud luua binariseeritud kaardi pealt, mille pealt tuvastus ka tehti, kuid see lähenemine on altim vigadele, kuna nimed ei eristu inimsilmalegi sama hästi kui värvilise kaardi pealt. Digitaalsel põhikaardil tuvastati 32 värvitooni, millest valged tekstid on kanalil 31. Sellele teadmisele toetudes on võimalik, et protsessi saab lihvida nii, et rõõpnimedest oleks midagi päästa, kuigi ka ülejäänud valged osad on samal kanalil. Testaladel moodustavad nad vähese osakaalu, välja arvatud testala 54082, millel esineb neid rohkesti. Tõhustamist sellest seisukohast võib kaaluda suurema andmemahu puhul.

Trükikaardilt mosaiikkaardilehe loomisel katsetati erinevaid masinnägemisele toetuvaid lähenemisi, mille keskmeks oli sisemise raamjoone tuvastamine (digitaalsel kaardil raami pole). Testaladel suurim viga raami ja kaardiala vahel on ≤ 22 m, mis on aktsepteeritav, sest trükikaardil talunimede väiketähtede kõrgus vastab < 25 meetrile. See aga ei tähenda, et vigase koha peale ei saaks sattuda nimi, mis siis saab puudulikult tuvastatud. Väiksema andmestiku puhul lahenduseks on GIS-tarkvaras kas joonistada ristkülikud käsitsi või tõsta need käsitsi paika. Suurema andmestiku puhul sunnib töös toodud lahendus vigadega leppima, ja lähenemine rõhutab järjepideva visuaalse kontrolli olulisust. Vea näidis on toodud joonisel 37.

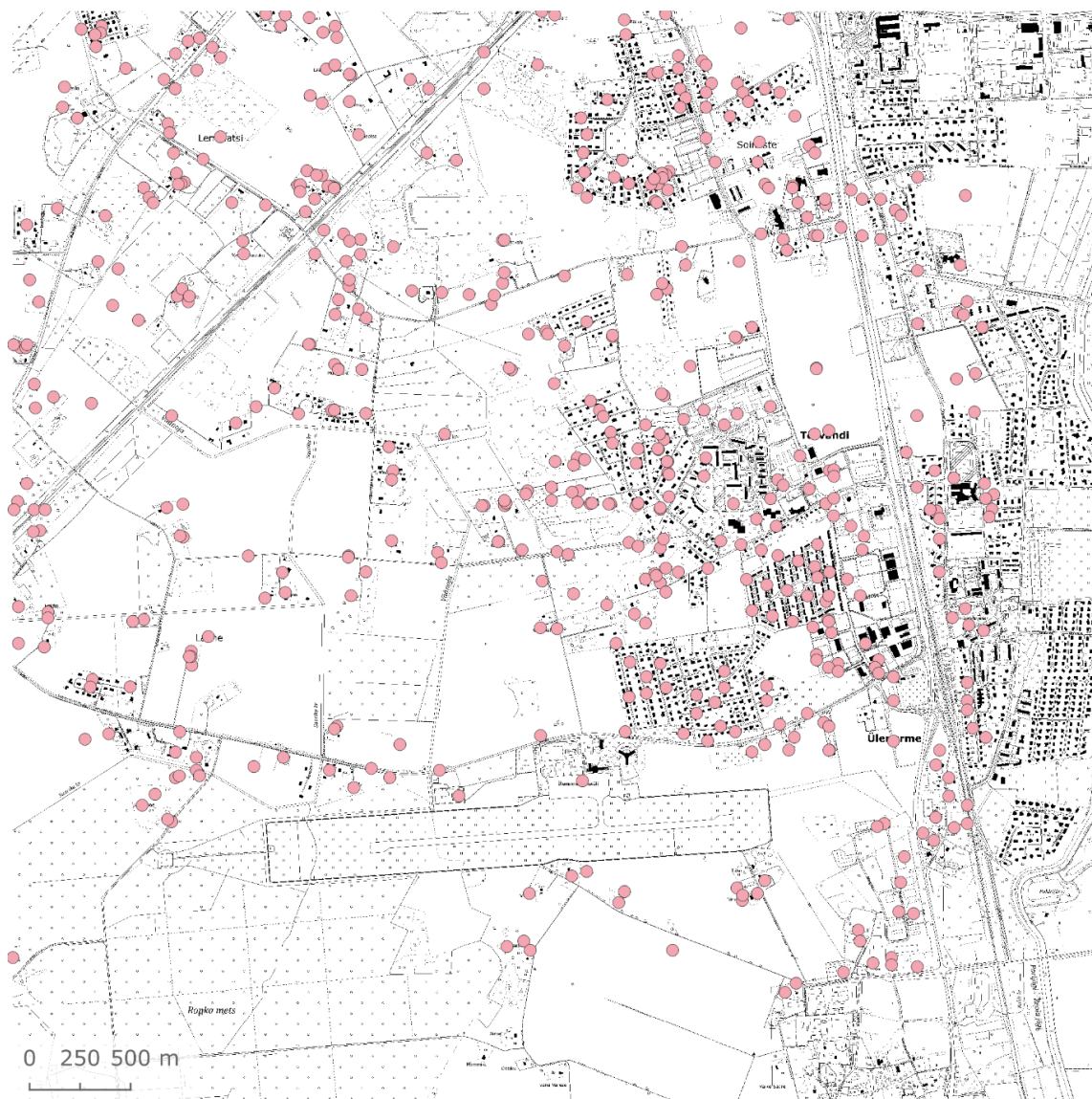


Joonis 37. Viga mosaiikkaardilehe loomisel toetudes masinnägemisele. Binariseerimata trükikaart, testala 54654. Kaardi allikas: Maa- ja Ruumiamet (2026).

Enamus mitmemodaalsete mudelite testidest tehti 6x6 ruudustikuga (valik on tulemuste tabelites lisades 6a ja 6b), kuid kohati ka väiksem võiks olla piisav. Näiteks Tartu ümbruse kehtiva põhikaardi peal (testala 54654) on rohkelt tihedat asustust, kus talunimesid ei esine, ja seega nende ruutude analüüsimine kulutab ressursse. Tasulised mudelid leiavad väiksemate ruutude puhul rohkem sõnesid – näiteks mainitud kohas on puhastatud tulemustes ehk sellistes, millel on väljundis koordinaadid ja mis ei koosne ainult arvudest kuid mille olemist testandmestiku risküliku sees pole kontrollitud, 3x3 ruudustiku puhul 98 ja 6x6 ruudustikus 524 sõnet. (Järgmine sisaldab siiski näiteks mitu korda sõne „õu“.) Ruudustiku suuruse küsimuses mõju võib avaldada ka see, milline maastik on – näiteks testalal 63644 pool lehest on soomaad, kus esineb ainult üksikuid nimesid. Lisaks eraldi teemana tuleks süüvida ressursside kasutamisesse ehk skripti optimeerimisse ja näiteks ruudustiku optimaalse suuruse leidmisse. Erineva koguse ruutudeks (3x3 või 6x6) jagamise mõjuga ei tehtud piisavalt teste, et tulemusi saaks võrrelda.

Keelemudeli tuvastuse jaoks tuleb leida üks näidisnimi. Näidisnime põhimõtte võiks siiski toimida paremini kaardil, kus ühte liiki nimed eristuvad teistest mitte ainult suuruse pärast, mis näidispildil ei väljendu. Kui asustusüksuste nimed on sama fondi, värvi ja sõrendusega kui talunimed, mudel leiab rohkem nimesid kui vaja, mis on ka ressursiküsimus.

Töövoo veaks võib lugeda selle, et tasulised mudelid jätsid tuvastamata nimesid, mis esinevad kaardilehe ida- või lõunaservas (joonis 38) ka siis, kui klasterdumise probleem (vt ptk 3.1.1 Sobivate mudelite tulemuste hindamine) oli ära lahendatud. Lähem uurimine tõi välja, et kaardi ruutudeks jagamise tulemusruudud ei ole ära katnud tervet kaardilehte. See on parandatud skriptisse, kuid töö maht ei lubanud tervet tuvastusprotsessi uuesti läbi viia, ning paranduse tulemusel jagatakse leht suuremaks arvuks ruute kui skriptis määratud. Nimede koonduvus avaldab mõju sellele, kui palju vea pärast nimesid tuvastamata jääb. Näiteks testalal 63644 (Kiili) nimed asuvad peamiselt kaardilehe põhjaosas ja välja jääb üksainus talunimi, kuid näiteks Võrumaal (testala 54082) on jaotus tasasem ja seega mõju suurem.

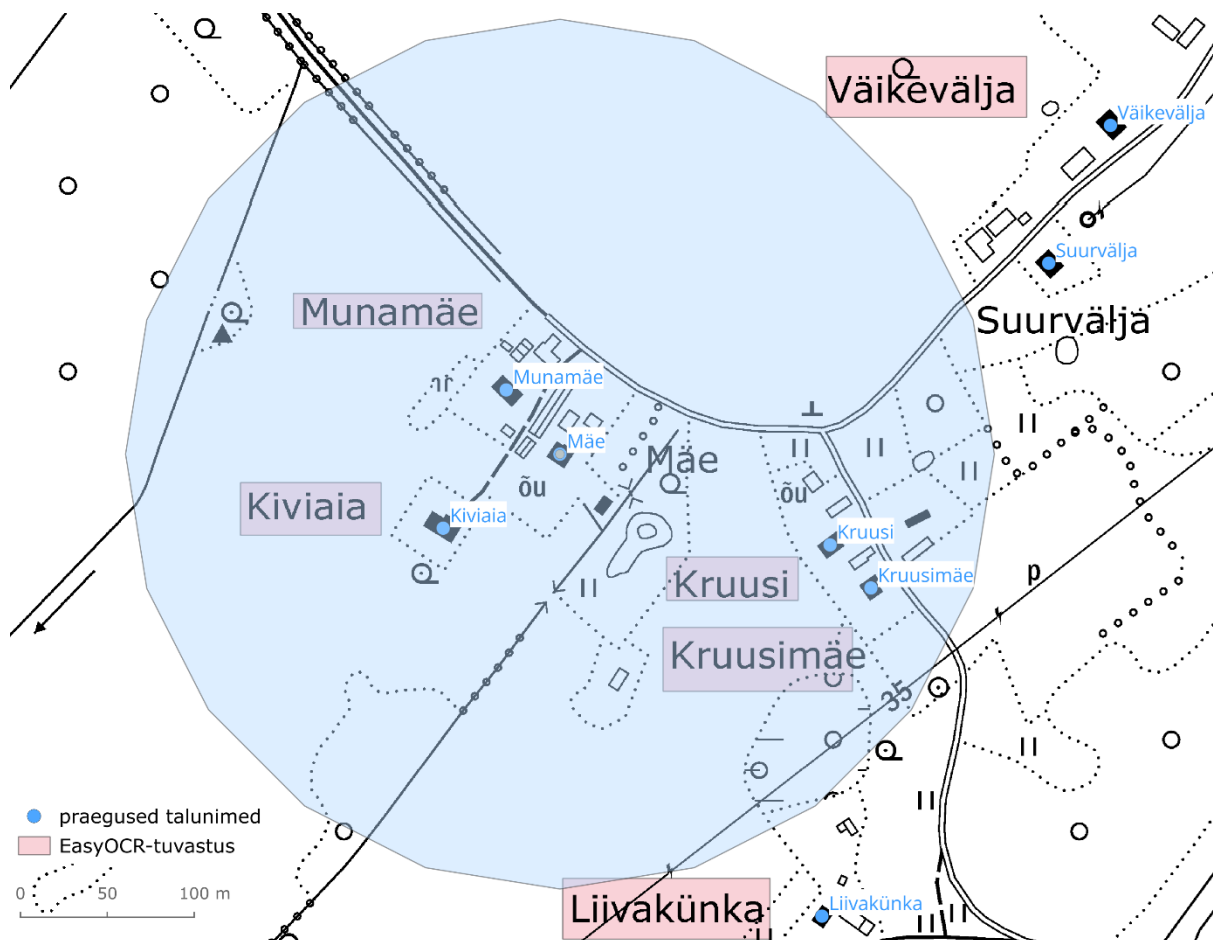


Joonis 38. Gemini 2.5 tuvastuspunktide jaotus kehtiva põhikaardi lehel, testala 54654. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

Andmete järeltöötluses oli vaja kokku liita kokku need nimeosad, mis kokku kuuluvad. Kombineeritud meetodis tingimuseks kehtestati, et liita nimed, kui riskülikute puhvrid laiuselt viis meetrit lõikuvad üksteisega ja kui üks või teine nimeosa sisaldab sidekriipsu ning et liitmine tuleb pärast kahe osa liitmist katkestada. Tingimuses lähtutakse sellest, et keelemudel tuvastab hargtäiendiga nimed tihti kahe osana ja ühes või teises osas sidekriipsu, mis võib osutuda suurema andmestiku puhul veaks, kuid ilma selle tingimusega liidetakse kokku kaks üksteisele lähestikku asuvat nime, mis on tegelikult eraldi nimed. See lähenemine ei toimi teistmoodi struktuuriga nimede tuvastamisel, sest näiteks „Vääna“ ja „jõgi“ jääksid liitmata, aga suure tähega talunimede jaoks on see enamasti tulemuslik. Siiski see ei lahenda olusid, kus liitsõna tuvastatakse kahe osana (näiteks *Pärn* ja *metsa*) või kui hargtäiendiga nime puhul pole tuvastatud sidekriipsu. Puudujääk avaldab mõju ka erinevate ajaliste kihistute kokkuviiimisel. Puhvri kitsendamine oleks lahendanud selle probleemi, ent jätnud paljusid nimesid kokkuliitmata, mida peeti suuremaks vigade allikaks. Tingimus säilitab sidekriipsu nii hargtäiendi lõpus kui ka põhiosa alguses juhul, kui see on mõlema osa puhul tuvastatud, kuid andmeid saab edasi regulaaravaldisega puhastada säilitades ainult ühe sidekriipsu. Samuti

ainuüksi keelemudeli alusel tehtud tuvastuses on võimalik, et informatsiooni on kadunud, kuna ülekattega pärast valiti kahest nimest tingimusele vastav, mitte ei sulandatud neid.

Et üle saada sellest, et teisenduskaugus on range mõõdik, mis ei tee vahet näiteks kahel järjestikul sidekriipsul ja erineval tähel, valiti erinevate aegade nimekihistute kokkuviimisel lisatingimuseks see, et sama teisenduskauguse puhul valitakse ruumiliselt lähem kandidaat. See aga ei päästa sellest, et kui kombineeritud meetodiga pole nime tuvastatud, pannakse sellele vale vaste. Näiteks EasyOCR pole leidnud nime *Mäe*, mis pärast selle praegune talunimi viiakse kokku nimega *Munamäe*, mille teisenduskaugus on puhvri sisse jäävatest kandidaatidest väikseim (joonis 39). Olekski hea leida mudel, mis saab leidmisega paremini hakkama. Sellise olemasolul on kombineeritud meetod adekvaatne lähenemine. Eeltoodud *Pärnametsa* näide kehtib ka nimekihistute kokkuviimise puhul ehk nimede sulandamise reegleid tuleks lihvida suurema andmestiku jaoks. Puhvri laiuse kasvatamine tekitab olukorra, kus kokku viiakse nimesid või nimeosasid, mis kokku ei kuulu, kuid lahenduseks võiks olla nimede joondamisega arvestamise lisamine tingimuseks. Alustades praeguste talunimedega võivad ka need raskusi tekitada. Neid filtreeriti regulaaravaldisega (vt ptk 2.1 Praeguste talunimedele leidmine), kuid andmestikus esineb ka rajatisi, mida avaldisega ei leitud. Sellised koosnevad ühest sõnast nagu *Pastoraadi* või *Vallamaja*, mis võivad omada vähest teisenduskaugust mõne teise nimega.



Joonis 39. Tuvastamata jäänud nime (Mäe) kokkuviimise variandid. Kehtiv põhikaart, testala 64593. Aluskaardi allikas: Maa- ja Ruumiamet (2026).

Üldistatud teisenduskaugus, milles muutustele defineeritakse kaalud (Orasmaa et al., 2010), ei päästa eeltoodud probleemidest, kui näiteks hargtäiendiga nime teist osa pole tuvastatud. Küll

aga võiks see aidata ajaliste kihistute kokkuviimisel, sest nimede arengule iseloomulik on see, et algne nimi on saanud juurde hargtäiendi, mis on kooskõlas ka Kohanimenõukogu (KNN, 1995) soovitustega. Lisaks keeleliste reeglite lisamine parandaks tulemuslikkust näiteks defineerides hargtäiendid elementideks, mis ei saa iseseisvalt esineda, kuigi praeguste talunimede andmestikus (mitte testaladel) esineb nii *Väike* kui ka *Vana*. Kui tahta kasutada tuvastamisel ainult EasyOCR-i, võiks sensitiivne teisenduskaugus (Davis & de Salles, 2007) olla ka lahendus, kuna EasyOCR tuvastab mõned tähed ja eriti täpitähed valesti. Samuti EasyOCR-i tulemuste puhastamises kasutatud regulaaravaldist tuleks lihvida, kuna mõned Võrumaa nimed sisaldavad ülakoma (näiteks „Hol’o“), ja hetkeseisuga see eemaldatakse.

Tulemuste arvulise hinnangu (F1-skoori) arvutamise aluseks võeti tuvastatud punktidest need, mis olid testandmestiku riskülikute sees, ehk need punktid loeti õigeteks positiivseteks. Vabavaraliste mudelite ja eriti EasyOCR-i puhul taoline lähenemine toimis – Paddle OCR-i problemaatikat on toodud varasemas peatükis –, sest see tagastab ristkoordinaadid korrektselt. Tasuliste mudelite puhul lähenemine muutis tulemust nõrgemaks, sest tulenevalt sellest, et koordinaatide viga oli pigem juhusliku iseloomuga, jäid mõned korrektselt tuvastatud nimed välja, kuna nad ei jäänud riskülikute sisse. Näiteks testala 54082 kehtival põhikaardil on kirja pandud 88 Gemini 3-ga tuvastatud nime 126-st, kuid 19 nime olid lisaks sellised, mille punkt asus risküliku vahetus läheduses, ja sama nähtus kordus ka teiste kaardilehtede peal. Tasuliste mudelite puhul tulekski edasisest valida teistmoodi hindamise alus ja kasutada näiteks puhvrit koos nimede lingvistilise võrdlemisega, kuid käesolevas töös piirduti ruumipäringuga, sest selle abil saadud suurusjärku peeti piisavaks hinnanguks ning sooviti kasutada hindamisel ühte ühtset meetodit.

Üldiselt soovituslik on visualiseerida vahetulemusi endale sobivas tarkvaras. Käesolevas töös on tegeletud viie testalaga ja kuigi paljud kaardid on standardiseeritud, ei saa välistada seda, et andmetes esineb iseärasusi. Lisaks kaasates ajaloolisi kaarte kõigub praeguste lehtede ulatus nende suhtes, kuna näiteks raamid pole alati sirged, ja seega vajalikke andmete allalaadimiseks oleks vaja saada GIS-tarkvaras näha, mis lehti testalaga kaetakse. Lisas toodud skriptis kõik vajalikud failid on alla laaditud arvuti kõvakettale, millelt neid loetakse. Tõhusam lahendus oleks aga lugemist teha serveris. Mõne koha peal saab skripti tõhusust tõsta, näiteks kirjutades mõned sammud funktsioonidena, kuid skript on praktiline väljund ja selgitus töö käigust, mitte valmis toode.

Arvestada tuleb ka aja ja ressurssidega, eriti kui kasutada suuremat andmestikku. Mitmemodaalsete mudelite arvutusaeg on minutites, kui kaardileht on jagatud ruutudeks ja alla minuti, kui tegemist on EasyOCR-i poolt leitud vähimate ümbritsevate riskülikute sees asuvate nimede tuvastamisega. Teiselt poolt PaddleOCR tuvastas CPU-ga ühte lehte tunni. Täpne aeg sõltub loomulikult arvuti jõudlusest ning keelemudelite puhul ka ühenduse kiirusest. Siiski võib-olla, et terve Eesti nimestiku leidmist ja lugemist pole mõistlik püüda korraga ära teha vaid näiteks kaardilehtede kaupa, või tuleks selleks skripti ümber kirjutada. Samuti nimede kokkupanekul pole välistatud, et suurema andmestiku puhul ei võiks välja tulla, et mingi teistmoodi lähenemine on eelistatavam.

Riskülikute kokkupanemiseks nende optimaalsete mõõtude arvutamine, nagu seda on teinud Schlegel (2021), talunimed oleks oma homogeensuses hea andmestik. Erinevate aegade kaartidel on need siiski erineva väljanägemisega, mispärast mõõdud tuleks arvutada iga kaardi jaoks eraldi, ning erinevad tuvastusmudelid võivad anda ka erinevate mõõtudega riskülikuid,

ehk mõõtude defineerimine oleks omaette teema ja sellepärast eelistati siintoodud lähenemist. Üks võimalus tulemusi parandada võiks siiski olla ristkülikute kokkupanek.

Skriptis on toodud ettetulnud probleemide lahendused. Eriti mitmemoodaalsete mudelite puhul on aga teatud etteennustamatust, kuna nende hingeelu ei tunta. See tähendab, et tuvastusprotsessi käigus võib tekkida probleeme ja veateateid, mida ei ole suudetud ette näha. Näiteks väljundfail on enamasti teatud kujul, aga kohati võib esineda ka teistmoodi formaati. Ettetulnud erandid on lahendatud ja tuvastustulemused ühesele kujule viidud, aga autorina ei saa välistada, et mudel ei saaks tekitada mingi kolmanda formaadi, mille normaliseerimise töökäiku pole olnud võimalik ette näha.

Teadusele iseloomuliku töö korratavuse seisukohast töös kasutatud kaardid tõenäoliselt ei tekita suurt vahelduvust. Testandmestik võib olla erineva suurusega tulenevalt koostajast – ka antud töös tuli välja, et saagis pole 100 % – kuid see ei mõjuta mudelite hindamist. Teisisõnu autori hinnangul kaartidel ei esine nimesid, mille puhul tekiks kahtlus, kas tegu on talunimega. Näiteks skeemilisel katastrikaardil omaette teemaks on asundustalud, mille tunnus koosneb tähest ja arvust, ja nende kaasamise osas tehtav otsus mõjutab oluliselt testandmestiku suurust. Korratavus on testitud vabavaraliste mudelite puhul hea, kuna need on pigem deterministlikud ehk tulemus on sama lähteandmestiku puhul pigem sama. Erinevusi võib tekkida ka kui näiteks puhver on natuke kitsam või laiem. Tasulised mudelid aga põhinevad tõenäosusele, mis tähendab, et tulemus saab olla erinev. Sättides temperatuuri nulli peale saab tõsta tõenäosust, et tulemus on sama, aga selles suhtes kindel olla ei saa. Töövoog tervikuna aga on korratav.

Long et al. (2021) kirjutavad sügavõppe rakendamise algusfaasi (2010ndad) kohta, et toonased töövood olid suurtele edusammudele vaatamata mitmefaasilised ja sellega aeglased ja keerulised. Sama kriitikat saab rakendada ka käesoleva töö väljundi kohta, kuna töövoog eeldab mitu sammu, kuid samas tegemist on esimese eestikeelsele kaardikogule kavandatud töövooga, mis arvestab just nende kaartide ja eesti keele eripäradega.

3.4 Edasised uurimissuunad ja järeldused

Edasistest uurimissuundadest kõige kesksem on vanemate kaartide kaasamine ja nende tuvastusega seotud probleemide lahendamine. Maa- ja Ruumiameti georefereeritud kaartidest sobivad talunimede esinemise poolest skeemiline katastrikaart (1930–1944) ning Nõukogude Liidu ja eestiaeegsed topograafilised kaardid. Katastrikaardil on toodud katastriüksuste nimed, kuid toonase käsitluse järgi langesid need kokku talunimega (välja arvatud see olukord, et käibel olnud nimi võis erineda ametlikust nimest nagu teoorias mainitud, vt nt Linder, 2007). Nõukogude Liidu topograafiline kaart O42 mõõtkavas 1:25 000 on oluline link esimese Eesti vabariigi ja taasiseseisvunud Eesti vahel ja kaardiseeria kõige vanem väljaanne peegeldab Eesti aja lõpu talude kihistut, kuid uuemates on juba Nõukogude erosiooni. Skeemiliselt katastrikaardilt on puudu Saaremaa ja mõned üksikud lehed ning eestiaeegsed topograafilised kaardid katvad ainult alla poole Eestist ära, 25-tuhandeline isegi pigem veerandi. Sellepärast tuleb vajadusel topograafilised kaardid kokku panna, kuigi mõõtkava on erinev, et saada Mandri-Eesti ära kaetud. Kaartide kattumisel tuleks eelistada suurema mõõtkavaga karti, sest väiksemas mõõtkavas hakkab generaliseerimise aste mõjutama nimede tihedust. Teisisõnu suurema mõõtkavaga kaardil on rohkem kohalikke nimesid, antud juhul talunimesid, ja väiksema mõõtkavaga kaardil on nendest näidatud ainult kõige tähtsamad, näiteks olulises

ristis asuvad talud. Lisaks võib kaalutleda näiteks sobiliku Saaremaa kaardi georefereerimist, et ruumilist alaesindatust tasandada. Ajalooliste kaartide puhul tuleb aga kohati leppida teatud lünklikkusega ruumilises kattes ja võimaluse korral saab kaasata ka teisi andmeallikaid. Lisaks on ka ajalooliste kaartide kvaliteet vahelduv.

Skeemilise katastrikaardi eripäradeks on käsitsi sisse viidud parandused, üle-eestiliselt vahelduv kirjalaad, piklikel ja mittehorisontaalsetel katastriüksustel pindobjekti suunda jälgiv nime paigutus, samasuguse kirja kasutamine ka teistel objektidel nagu viljapuuad (mida esineb aga ka eestiaegsel kaardil ja eristus tuleb teha punktlemmärgi alusel) ja kõikuv kirjaviis ehk näiteks kõrvuti asuvate üksuste nimed võivad kajastada kahte erinevat õigekirjaviisi ja pealegi omada sama arvulise osa. Lisaks kohati esineb üksteise peale trükitud nimesid. Kõik see seab nõudmisi tuvastusmudeli valiku suhtes, ent tehnoloogia kiire areng annab lootust, et need pole ületamatud probleemid. Abi võiks saada ka katastriüksuste pindade automatiseeritud vektordamisest, ja kaartidelt elementide tuvastamine ongi tõusev uurimissuund.

Talunimede säilimise analüüsi saaks edasiselt täiendada kaasates tänavanimestiku, milles sageli säilitatakse vanemat nimekihti. Näiteks Tartu ümbruse trükikaardi lääneserva talunimi *Lepalaane* on praeguseks *Lepalaane tee*, *Juuristiku* asemel on Laane külas *Juuristiku tee* ja *Künnapuu* läheduses Tõrvandis on *Künnapuu tee*. Tänavanimestiku motiveeritus nendes kohtades on ilmne, sest kokkusattumusteks on näidiseid liialt ja nimestik pole ka kõige sagedasem. Lisaks huvitav oleks uurida talude asukohtade ajaloolisi muutusi kasutades hoonete ja katastriüksuste pindobjekte. Nime asukoht annab teabe talu asukoha kohta ainult umbkaudselt, ainult niivõrd, mida on vaja selleks, et kaardi kasutaja saab nime ja selle viidatud objekti kokku viia. Nime paigutus järgib teatuid hierarhilisi reegleid (vt Suurna & Sisas, 2010), kuid ruumipuudusel tuleb nimi panna sinna, kuhu see mahub (vt Mõisja, 2019). Sellepärast puhtalt nimede analüüsiga ei saa vastatud sellele, missuguseid muutuseid talude asukohtades on toimunud. Taolisse analüüsi tasuks lisada ka teisi ruumilise info kihte, et kaasata maakasutuse muutused. Selleks masinõppe ja eelmainitud automatiseeritud vektordamine pakkuvad häid võimalusi.

Üks järgnev samm olekski koostada talunimede andmebaas, mis hõlmaks iga nime ajaloolise arengu. See aga eeldab nimekihistute kokkuviiamise keeletehnoloogilist lahendust. Algust saab teha põhikaartide alusel käesoleva töö tulemustel ja hiljem laiendada ajalooliste kaartide peale. Suuremahulisem kartograafiline andmebaas annabki võimaluse seoseid leida (Saar, 2008a). Kaartide valiku ja andmebaasi hilisem täiendamine teiste allikatega ongi osa teaduse ennast parandavast lähenemisest, sest teaduslikku uurimust iseloomustab see, et üksiku uurimuse raames ei saa uurimisvaldkond valmis.

Kuna mitmemodaalsetele mudelite tuleb ette anda viip (prompt), võiks katsetada selle sõnastuse mõju tulemusele. Antud töös mudelitele on kirjutatud, et kaart on eesti keeles, aga vastuseta jääb, kas eestikeelse viibaga tulemus oleks olnud teistsugune. Teine sõnastuse aspekt on tagastatav geomeetria. Ristkülik oleks edasise töötamise seisukohast kergem ja geomeetriliselt kindlam. Käesoleva töö väljundiks on koordinaatpaariga punkt ja punktidega kaasnes vigu, kuid ristkülikutega tulemus oli veelgi kehvem. Tekste ümbritsevate ristkülikute joonistamine ei ole mitmemodaalsete mudelite pärusmaa ja nad ei pruugi sellega iga kord hästi hakkama saada. Kui tulemus üldse on rahuldav, koordinaadid võivad olla formaatides, mis nõuavad järeltöötlust. Selle probleemi kõrvaldamine lihtsustaks oluliselt tuvastust ja tõstaks saagist, kuid pole välistatud, et sellel puhul ainuüksi sõnastusest ei piisa.

Seisuga 1/26 on Eesti Keele Instituudis käsil projekt, milles luuakse edetabel, milles erinevad tehisaru mudelid on järjestatud pingeritta faktiteadmiste, valeinforiski, Eesti õigusruumi tundmise ja eesti keele oskuse alusel (Kruusmaa, 2026). Taoline tabel annab ülevaate mudelite võimekusest erinevates valdkondades ja seega lihtsustab sobiliku mudeli valikut vastavalt ülesandele (Kruusmaa, 2026). Kuigi tabeli kasutusala võib pigem poliitiliste otsuste tegemisel marjaks ära kuluda, ei saa eita selle tähtsust ka edasisel kartograafilisel analüüsil, kus tähtis on eesti keele oskus – vastupidiseks näiteks olgu ka käesolevas töös välja toodud MMOCR-i ingliskeelsed tulemused eestikeelselt kaardilt – ning tähtis teadmine on kindlasti ka arusaam Eesti kartograafilistest konventsioonidest, mis väljendavad kultuuri hinnanguid.

Töö tulemustel saab öelda, et kuigi tuvastusmudelid on arenenud sügavõppe kasutuselevõttuga, kaardid on endiselt väljakutseid pakkuv materjal ja paljud tuvastusmudelid jäävad nendega hätta. Kaartidel esineb nimesid, mis on ka inimsilmale raskesti tuvastatavad, seega kui neid ei otsita näiteks tooni alusel – mis võiks olla edasiarendus – vähemalt tänase tehnika puhul jääb osa tuvastamata. Tuvastamata jäänud nimede kohta ei leitud tendentsi lisaks kaardi mürale. Mõne puhul seletuseks on asukoht hoonestuse, nimede ja/või punktleppemärkide poolest tihedal alal, kuid see ei kehti kõikide puhul. Nendest mudelitest, mille peale ehitati lõplik töövoog, parimad tulemused olid kombineeritud meetodil, mis ühendab hea geomeetrilise täpsusega EasyOCR-i ja mitmemodaalse mudeli „keeleoskuse“ tugevused, kuigi mõned nimed leiti puudulikult. Eraldi nende nõrkusteks on EasyOCR-il tuvastuse täpsus (ebakorrektsus) ja mitmemodaalsetel mudelitel koordinaadivead. EasyOCR sobib aga ainsa meetodina kaardilehtedele, mis pole informatsiooni poolest sisutihedad ja milles talunimed eristuvad hästi. Sellel puhul tähtis on aga andmete puhastamine regulaaravaldisega. Puudujääkidega tuleb sõltumatult meetodist siiski leppida ja ka adekvaatseks hinnatud EasyOCR-i puhastatud tulemuste 17 testist ainult neli jäi tavapärase viie protsendi veamarginaali ($F1\text{-skoor} \geq 0,95$) sisse. Kombineeritud meetodis puhastamine polnud vajalik, kuid sellele vaatamata kolmandik skooridest ületas 0,95 piiri. Töö kirjutamise aeg on aga tehisaru võimaluste suhtes murraku aeg. Hetkeseisuga mitmemodaalsete mudelite asukohatäpsus on liialt nihkes, et pakkuda tuvastamiseks adekvaatset alternatiivi, kui nimed on vaja kaardile ka saada. Testimisel oli aga täheldada, et Gemini 3 andis süstemaatilisema koordinaatide vea kui Gemini 2.5, ehk kui korrata töövoogu mõne aja pärast ka sama mudeli uuema versiooniga, tulemus võib olla oluliselt parem. (Mudeleid arendatakse teatud huvidest lähtuvalt ja seega neid treenitakse näiteks suuremate keeltega ja kõik ei toeta eesti keeltki, ent siiski ei tekki taolist probleemi, et inimlugeja näiteks üldse vene keelt ei valda.)

Kokkuvõte

Kohanimed omavad tähtsust kartograafidele, planeerijatele ja omavalitsustele, kuid need pakkuvad ka vaadet ajalukku, mis omapoolt on mineviku uurijate jaoks väärtuslik. Talunimedes on säilinud vanemaid ajalisi kihistuid ja vanemad talunimed omapoolt esinevad näiteks tänavanimedes. Sellele vaatamata on talunimedele pööratud üsna vähe tähelepanu ja näiteks terviklikku Eesti talunimede loendit autori teada pole olemas.

Käesolev magistritöö on keskendunud Eesti talunimede ja töö eesmärgiks on olnud leida sobilik lähenemine talunimede automatiseeritud tuvastamiseks Eesti põhikaartidelt ja selle alusel luua praktiline väljund ehk Python-skript, mis hõlmab terve tuvastusprotsessi. Eesmärgi saavutamiseks on võrreldud erinevate tekstituvastusmudelite (OCR, *Optical Character Recognition*) võimekust ülesande jaoks. Mudelid olid EasyOCR, MMOCR, NewOCR, PaddleOCR, Tesseract, kolm mitmemodaalset mudelit Gemini 2.5 Pro, Gemini 3 Flash ja Qwen3-VL-32B ning kombineeritud meetod, milles tekstid leitakse EasyOCR-i ja tuvastatakse Gemini 3-ega. Mudelite katsetamiseks on kasutatud viite testala, mis vastavad kehtiva põhikaardi 1:10 000 lehele (5x5 km) koos 200-meetrilise puhvriga kaasamaks ka servas lõikuvad nimed. Testalad püüti valida võimalikult mitmekesisteks arvestades keelt, nimede ruumilist jaotust ja kaartide väljanägemist. Lisaks on loodud testandmestik, millesse sisalduvad käsitsi korjatud talunimed nende alade kolmelt erineva aja põhikaardilt ajavahemikus 1996–2026 (kaks digitaalkaarti 1:10 000 ning trükikaart 1:20 000). Testandmetikus on kokku 1084 nime ja neid kasutati mudelite hindamises nii-öelda põhitöena. Hindamise mõõdikuks kasutati F1-skoori, mis on saagise ja täpsuse harmooniline keskmine. Saagis mõõdab seda, kui palju nimesid tuvastusega korrektselt kätte saadakse osakaaluna kõikidest kaardil esinevatest nimedest ja täpsus mõõdab, kui suur osakaal tuvastatud sõnedest on õigesti tuvastatud.

Töö tulemustel saab öelda, et eestikeelsele ja murdelisi jooni sisaldavale talunimede andmestikule sobivad EasyOCR, mitmemodaalsed mudelid ning kombineeritud meetod. Väljundskriptis on toodud töövoog kõikide nende jaoks eraldi. Esimese puhul nõrkuseks on siiski keerulisematel kaartidel nimede puudulik leidmine, mis kajastub ka kombineeritud meetodi tulemustes, ning mitmemodaalsete mudelite puhul see, et nende raskused ruumilise täpsuse mõistmisel viivad juba piksliruumis valede asukohtadeni, mis pärast saagis jääb tegelikult madalamaks, kuna koordinaatide nihe ei ole süstemaatiline. Tulemusi saakski parandada valides leidmiseks mingi teise mudeli EasyOCR-i asemel, kuid selle töö raames sobilikku ei leitud, ning EasyOCR-i kasuks räägib selle lihtsus ja eeskätt võime säilitada kohateave ehk väljundiga on kerge GIS-tarkvaras edasi tegeleda. Mitmemodaalsete mudelite tugevus seisneb nende kõrges tuvastustäpsuses.

Eriti keerulisteks osutusid palju samakõrgusjooni sisaldavad Võrumaa kaardid, mille puhul tuvastustulemused kannatasid müra all. Edasiseks tulekski katsetada binaarkaartide teistmoodi tekitamist, kuid samas tehisaru võimaluste areng on niivõrd kiire, et kui nende ruumiline täpsus saadakse paremaks, ka värvilised kaardid sobivad sisenditena. EasyOCR sobib ainsa lahendusena kaartidele, millel müra eriti ei esine. Ülejäänud testitud mudelite väljundites esines müra, mida ei saa regulaaravaldisega tõhusalt eemaldada, treenimisandmetest tulenevaid inglisekeelseid tuvastusi ning pigem nõrkasid tulemusi koos vigaste koordinaatidega. Need sobivadki pigem OCR-i algse eesmärgi jaoks ehk sirgerealise ja puhta taustaga teksti tuvastamiseks. Arvuliselt EasyOCR-i puhastatud tulemuste 17 testist neli jäi tavapärase viie

protsendi veamarginaali ($F1\text{-skoor} \geq 0,95$) sisse ja kombineeritud meetodis kolmandik skooridest ületas piiri. Ka varasemate töödega võrreldes (nt Dörterler et al., 2025) on tulemused enamasti heal tasemel. Eestikeelne andmestik sobis nendele mudelitele, mille peale lõplik töövoog ehitati, ja murdelised jooned ei avaldanud tulemustes mõju.

Erinevaid ajalisi nimekihte viidi kokku teisenduskauguse abil saamaks teavet selle kohta, mis nimed on püsinud, kuigi sellega saadud hulgad ei tarvitse ühtida. Teisenduskaugus tähendab, et loetakse kokku, kui mitu teisendust tuleb teha saamaks ühest sõnest teine (Jaanimäe, 2021). ehk hinnatakse sõnade omavahelist sarnasust. Üldistatud teisenduskauguse abil saab teisendustele defineerida kaalud ja sellega saaks tulemusi täpsustada, ent tegelikud nimemuutused on niivõrd mitmekesised, et kõikide nimede kokkuviiamiseks on vaja lisada teisi lähenemisi. Levenshteini teisenduskauguse kasutamine aitab siiski valgustada esialgseid probleeme nimekihistute ühendamisel.

Töö uudsus seisneb selles, et autori teada esimest korda on katsetatud ja hinnatud erinevaid tekstituvastusmudeleid Eesti kaardiandmestiku peal. Ei eesti keel ega kaardid pole peavoolu OCR lahenduste peamises fookuses ning nende tööriistade jõudlus selles väga spetsiifilises rakenduses pole garanteeritud – probleeme esineb nii täpitähtede tuvastuse kui teiste kaardil esinevate joonte ja sümbolitega ümbritsetud nimede piiride mõistmisega. Töö tulemused annavad ka esimest teavet talunimede nimekujude muutumisest ja muutuste uurimise võimalustest keeletehnoloogia lahendustega. Lisaks on loodud Eesti kaardiandmestiku jaoks kohandatud töövoog, ja ka selle jaoks vajalik andmete ettevalmistamine on oluline samm edasiste tuvastuse rakenduste seisukohast.

Automated recognition of farmstead names on Estonian Basic Maps

Maria Anttila

Summary

Place names provide a valuable insight into history and serve as an essential source for urban planners. Farmstead names in Estonia are also present in the newer layers of toponyms, for example in street names. Nevertheless, there is no known gazetteer for Estonian farmstead names. Current names of cadastral parcels, which also serve as farmstead names on dispersed settlement areas, can be obtained through a spatial query, but older names have traditionally been manually collected from maps and other sources, such as revision lists or contract books.

With the rise of machine learning and its applications in Optical Character Recognition (OCR), this process can be automated and further made more precise with the help of neural networks. Automated text recognition generally consists of two phases: detecting text boxes and recognizing their content (Wang et al., 2023). Classical OCR methods compare neighbouring pixels to extract text areas (Lin et al., 2020). Recently, large language models (LLM) together with large vision models (LVM) form vision-language models (VLM) and multimodal large language models (MLLM), which combine the reasoning performance of the language models and the clear sight of the vision models (Yin et al., 2024, see also Bordes et al., 2024; Tong et al., 2024). This development has significantly influenced OCR as newer models are able to see linguistic patterns and extract information based on context (Zhou, 2026). With regards to maps, it is more appropriate to speak about scene texts, as classical OCR is designed for text documents with a clear structure where text forms straight lines (see Milleville et al., 2022).

There is currently no known OCR pipeline for Estonian map data that would consider the specific nature of local maps or cover the complete workflow with the necessary pre- and post-processing. This thesis aims to address that gap by comparing various text recognition models, both freeware and paid ones, and evaluating their usability for Estonian-language map data. As the training data of machine learning models favour most widely spoken languages, it is not clear if they perform adequately with Estonian content, especially as the farmstead names have some dialectal features in them. The research material covers three editions of the Estonian Basic Map, which serve as a sample material for future research involving historical maps.

The research questions of the thesis are:

Is automated text recognition suitable for Estonian-language farmstead names containing stratified and dialectal elements?

What are the limits of the tested models' usability in automated text recognition on Estonian-language maps?

How does edit distance function in reconciling different temporal name layers?

The practical output of the thesis is a Python script covering the complete recognition process on an example of one map sheet. Besides the technical development of OCR as a method and the specification of maps as the source data of the recognition, the thesis also gives an overview on the previous research on Estonian farmstead names. The research is carried out on five test sites, which were chosen to be as heterogenous as possible in terms of language, spatial distribution of the names and map quality. Each site corresponds to one map sheet of the current digital 1:10,000 Estonian Basic Map, a 5x5 km area with an additional 200 m buffer to include

names cut off at the sheet edges. A test data of total 1,084 names was manually collected from all tested sheets and used to evaluate the models.

The compared approaches include five freeware models (EasyOCR, MMOCR, NewOCR, PaddleOCR and Tesseract), three multimodal models (Gemini 2.5 Pro, Gemini 3 Flash and Qwen3-VL-32B) and a combined method where text boxes are detected with EasyOCR and text is read using Gemini 3. The results revealed both model-specific and categorical problems. EasyOCR, all the multimodal models and the combined method proved suitable for the workflow. None demonstrated any notable difficulty with the dialectal features.

EasyOCR returns the coordinates correctly, which is a significant advantage on recognition on maps. Although EasyOCR supports Estonian, the results still include noise and occasional characters not belonging to the Estonian alphabet; however, an appropriate cleaned result was achieved with a simple regular expression. Despite the recent development, multimodal models still show shortcomings in spatial reasoning (see Liu et al., 2026), which resulted in lower recall, as noticed when converting pixel coordinates to the Estonian coordinate system. The precision of the multimodal models is mostly acceptable (7 out of 9 recognitions achieved a precision $\geq 0,95$ with cleaned data and edit distance of 0 or 1). The choice of model was inconsequential, but Gemini 3 performed slightly better on spelling than Qwen.

The combined method harnesses the strengths of the two approaches (the coordinate precision of EasyOCR and the recognition precision of the multimodal models) but its weakness stems from the partly poor text detection of EasyOCR. F1 score, i.e. the harmonic mean of precision and recall, was higher on map areas with less noise. One particularly challenging area was the rugged landscape of Võrumaa in South-Eastern Estonia. EasyOCR alone was sufficient on areas where maps are clear and without noise. No other tendency was seen in the undetected names for any of the models. Based on the results, for more complex maps, it is recommended to perform the recognition workflow with the combined method.

The other tested models suffered from noise that cannot be easily eliminated by rule-based processing, provided English outputs based on the language of their training data or combined a weak recognition result with a coordinate problem.

Different thresholds of edit distance were tested for the purpose of reconciling temporal name layers. This was done by comparing the current names of cadastral parcels with each recognized layer and by comparing each layer with the previous one. The first approach in particular can give information about names preserved across all layers, although the elements of the sets don't necessarily coincide. Generalized edit distance allows to define weights for changes, which could solve the theme of affixes, but the changes in the names are too diverse for edit distance alone to fully satisfy the reconciliation.

Nevertheless, there are certain bottlenecks in the workflow. The recall could likely be increased through colour-based binarization by distinguishing the black colour from the maps, which would eliminate the contour lines that interfere with the recognition. Additionally, the tiling was observed to be partly incomplete and there are areas in the eastern and southern edges of the sheets where the multimodal models did not perform recognition. The distribution of names affected the quantity of names that remained undetected due to the error. The issue has been corrected in the final version of the script that now, for the first time, provides a complete workflow for the recognition of Estonian farmstead names.

Tänuavaldused

Suurimad tänud kuuluvad töö juhendajatele Raivo Aunapile, Siim Orasmaale ja Ardi Tampuule, kes olid igakülselt abiks ja toeks – eriti suur tänu huvitava teema väljakäimise, heade ja edasiviivate mõtete ning tehnilise toe eest, mis aitas mind töö jooksul areneda. Täna ka kõiki oma kaasaelanud lähedasi.

Kasutatud kirjandus

- Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (lk 9357–9366). <https://doi.org/10.1109/CVPR.2019.00959>
- Bahjat, A. K. (2025). Automated document processing system for government agencies using DBNET++ and BART models. *International Journal of Circuit, Computing and Networking*, 6(2), 34–41. <https://doi.org/10.33545/27075923.2025.v6.i2a.100>
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv*. <https://doi.org/10.48550/arXiv.2308.12966>
- Bathae, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, 31(2), 889–938.
- Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H., & Chandra, V. (2024). An introduction to vision-language modeling. *arXiv*. <https://doi.org/10.48550/arXiv.2405.17247>
- Chen, X., Jin, L., Zhu, Y., Luo, C., & Wang, T. (2020). Text recognition in the wild: a survey. *ACM Computing Surveys (CSUR)*, 54(2), 1–35. <https://doi.org/10.1145/3440756>
- Chiang, Y.-Y., & Knoblock, C. A. (2015). Recognizing text in raster maps. *Geoinformatica*, 19, 1–27. <https://doi.org/10.1007/s10707-014-0203-9>
- Chiang, Y.-Y., Leyk, S., Honarvar Nazari, N., Moghaddam, S., & Tan, T. X. (2016). Assessing the impact of graphical quality on automatic text recognition in digital maps. *Computers & Geosciences*, 93, 21–35. <https://doi.org/10.1016/j.cageo.2016.04.013>
- Chiang, Y.-Y., Moghaddam, S., Gupta, S., Fernandes, R., & Knoblock, C. A. (2014). From map images to geographic names. *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14)*, 581–584. Association for Computing Machinery. <https://doi.org/10.1145/2666310.2666374>
- Davis, C. A. Jr., & de Salles, E. (2007). Approximate string matching for geographic names and personal names. *Proceedings of the IX Brazilian Symposium on Geoinformatics*, 49–60.
- de Mello, C. A. B., & Lins, R. D. (1999). A comparative study on OCR tools. *Proceedings of Vision Interface '99*, 224–232.
- Dörterler, S., Şahin, E., & Özdemirc, D. (2025). Evaluating scene text recognition models through multidimensional error analysis: Insights from Turkish scene text data. *Gazi Journal of Engineering Sciences*, 11(3), 495–521. <https://doi.org/10.30855/gmbd.070525AR09>
- Eesti Kirjandusmuuseum. (2024). *Erinevate tekstivastusmudelite tuvastustäpsuse hindamine Levenshteini meetodil*. Vaadatud 18.12.2025 <https://www.kirmus.ee/et/tegevus/tekstivastus>
- [EKI IT] = IT terministandardi sõnastik (2026). Vaadatud 9.2.2026 <https://arhiiv.eki.ee/dict/its/index.cgi?Q=rekurrentne>

- EKI ühendsõnastik (2026). Vaadatud 16.2.2026
<https://sonaveeb.ee/search/unif/dlall/dsall/t%C3%A4rktuvastus/1/est>
- [EKSS] = Eesti keele seletav sõnaraamat (2009). Vaadatud 11.3.2025
<https://arhiiv.eki.ee/dict/ekss/index.cgi?Q=talu>
- Ernits, E. (2011). Nõo kihelkonna talunimedest XIX sajandi keskpaiku. S. Jüva (toim), *Ruum, kotus ja kotussõnimeq* (lk 46–76). Võro Instituudi toimõndusõq 25. Võro Instituut.
- Faster, M. (2013). Nulgast mäeni. Ülevaade Setomaa nimevarast. A. Kiristaja. *Setomaa kohanimed* (lk 201–215). Seto instituudi toimetised 1. SA Seto Instituut.
- Google Lens (2026). Vaadatud 17.2.2026 <https://lens.google/>
- Haack, C. (2018). *A contribution to computer-assisted deciphering of labels from scanned maps*. [Master's thesis, Joined Cartography Master Program by TU Munich, TU Vienna, TU Dresden and University of Twente.]
- Hand, D. J., Christen, P., & Kirielle, N. (2021). F*: an interpretable transformation of the F-measure. *Machine Learning*, 110, 451–456. <https://doi.org/10.1007/s10994-021-05964-1>
- Harvalik, M. (2004). Hofnamen. A. Brendler, & S. Brendler (toim), *Namenarten und ihre Erforschung. Ein Lehrbuch für das Studium der Onomastik* (lk 415–425). Baar.
- Hernández-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13(1), 2813–2869.
- Jaanimäe, G. (2021). Ajalooliste tekstide normaliseerimine. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 17, 47–59. <https://doi.org/10.5128/ERYa17.03>
- Johansen, P. (1933). *Die Estlandliste des Liber Census Daniae. I. Halbband*. Estländische Druckerei Aktien-Gesellschaft.
- Jordan, P. (2009). Some considerations on the function of place names on maps. *Proceedings of the 24th International Cartographic Conference*. Vaadatud 10.10.2025
https://icaci.org/files/documents/ICC_proceedings/ICC2009/html/nonref/12_2.pdf
- Kallasmaa, M. (2023). Märkusi talunimede vanusest Audru näitel. *Keel ja kirjandus*, 2023(6), 595–604. <https://doi.org/10.54013/kk786a3>
- Kallasmaa, M. (2005). Eesti kohanimede liigitamisest. *Keel ja kirjandus*, 2005 (2), 136–140.
- Kallasmaa, M. (2001). Eesti talunimed ja germaani isikunimede probleem. T. Seilenthal (toim), *CONGRESSUS NONUS INTERNATIONALIS FENNO-UGRISTARUM 7.–13.8.2000 Tartu. Pars V. Dissertationes sectionum: Linguistica II* (lk 30–34). Trükk OÜ PAAR.
- Kallasmaa, M. (2000). *Saaremaa kohanimed II*. Eesti Keele Instituut.
- Kallasmaa, M. (1996). *Saaremaa kohanimed I*. Eesti Keele Instituut.
- Kallasmaa, M. (1995). Name Studies in Estonia. E. Eichler, G. Hilty, H. Löffler, H. Steger, & L. Zgusta (toim), *Namenforschung – Ein international Handbuch zur Onomastik*, 1. Teilband (lk 178–182). Walter de Gruyter.
- Kalle, I., Pikk, K., & Kurvits, Ü. (2006). *Kiidjärve mõis ja vanad talukohad: Kiidjärve koduloo-uuring*. Kiidjärve.

- Kepsu, S. (1987). Talonnimien tutkimisesta. E. Kiviniemi, & R. L. Pitkänen (toim), *Kieli 2 – Näkymiä nimistöön* (lk 35–73). Helsingin yliopiston suomen kielen laitos.
- Kim, J., Li, Z., Lin, Y., Namgung, M., Jang, L., & Chiang, Y.-Y. (2023). The mapKurator system: a complete pipeline for extracting and linking text from historical maps. *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23)*, 1–4. Association for Computing Machinery.
<https://doi.org/10.1145/3589132.3625579>
- Kiristaja, A. (2013). *Setomaa kohanimed*. Seto Instituudi toimetised 1. SA Seto Instituut.
- Klijn, E. (2008). The current state-of-art in newspaper digitization – a market perspective. *D-Lib Magazine*, 14(1/2). <https://www.dlib.org/dlib/january08/klijn/01klijn.html>
- [KNN] = Kohanimenõukogu (1995). *Vabariigi Valitsuse kohanimekomisjoni pöördumise kohalike omavalitsuste poole*. Vaadatud 2.10.2025 https://arhiiv.eki.ee/knn/knk_talu.htm
- [KNR] = Kohanimeraamat (2017). Vaadatud 10.10.2025
<https://arhiiv.eki.ee/dict/knr/index.cgi?Q=Nutru>
- [KNS] = Kohanimeseadus (2003). Vaadatud 25.1.2026 <https://www.riigiteataja.ee/akt/KNS>
- Kruusmaa, K. (2026, 21. jaanuar). *Krister Kruusmaa: kes valis ChatGPT riigikokku? Arvamus*. ERR. Vaadatud 26.1.2026 <https://www.err.ee/1609917185/krister-kruusmaa-kes-valis-chatgpt-riigikokku>
- Laansalu, T. (2014). Kose kihelkonna põliste asustusnimede vanus ja päritolu. *Emakeele seltsi aastaraamat*, 60, 101–126.
- Leht, M. Isiklik konsultatsioon e-posti teel 6.3.2025.
- Lenc, L., Martínek, J., Baloun, J., Prantl, M., & Král, P. (2022). Historical map toponym extraction for efficient information retrieval. S. Uchida, E. Barney, & V. Eglin (toim), *Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science: Vol. 13237*, (lk 171–183). Springer. https://doi.org/10.1007/978-3-031-06555-2_12
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.
- Li, H., Wang, P., Shen, C., & Zhang, G. (2019). Show, attend and read: a simple and strong baseline for irregular text recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 8610–8617. <https://doi.org/10.1609/aaai.v33i01.33018610>
- Lin, H., Yang, P., & Zhang, F. (2020). Review of scene text detection and recognition. *Archives of Computational Methods in Engineering*, 27, 433–454.
<https://doi.org/10.1007/s11831-019-09315-1>
- Linder, H. (koost). (2007). *Kaiu talud ja saunad. Mahtra Talurahvamuuseumi toimetised III*. Mahtra Talurahvamuuseum.
- Liu, D., Liang, T., Hu, Z., Peng, J., Lu, Y., Xu, Y., Fu, Y., & Yin, Y. (2026). Spatial intelligence in vision-language models: a comprehensive survey. Eeltrükk.
<https://doi.org/10.21203/rs.3.rs-8919250/v1>
- Liu, Z., Song, R., Li, K., & Li, Y. (2025). From detection to understanding: a systematic survey of deep learning for scene text processing. *Applied Sciences*, 15(17), 9247.
<https://doi.org/10.3390/app15179247>

- Long, S., He, X., & Yao, C. (2021). Scene text detection and recognition: the deep learning era. *International Journal of Computer Vision*, 129, 161–184. <https://doi.org/10.1007/s11263-020-01369-0>
- Maa-amet. (2022). *Aadressiandmete käsiraamat*. Vaadatud 7.4.2026 https://geoportaal.maaamet.ee/docs/aadress/aadressiandmete_kasiraamat.pdf
- Maa- ja Ruumiamet. (2025a). *Eesti põhikaardi 1:10 00 leppemärgid ; Legend of Estonian basic map 1:10 000*. Vaadatud 13.4.2026 https://geoportaal.maaamet.ee/docs/pohikaart/pk_legend.png
- Maa- ja Ruumiamet. (2025b). *Põhikaardistuse ajalugu*. Vaadatud 16.9.2025 <https://geoportaal.maaamet.ee/est/Andmed-ja-kaardid/Topograafilised-andmed/Eesti-Pohikaart-110-000/Pohikaardistuse-ajalugu-p113.html>
- Maa- ja Ruumiamet. (2024). *Kui kaardil on talu juures vale nimi, siis kuidas seda muuta?* Vaadatud 2.10.2025 <https://geoportaal.maaamet.ee/est/abi-ja-juhised/andmed/kui-kaardil-on-talu-juures-vale-nimi-siis-kuidas-seda-muuta/84/>
- Mane, A., Satape, M., Kotecha, K., Bagul, A., & Chavan, S. (2025). Image format conversion methods for digital media. *International Journal of Research Publication and Reviews*, 6(3), 153–160.
- mapKurator. (2023). *Ongoing Research*. Vaadatud 12.2.2026 <https://knowledge-computing.github.io/mapkurator-doc/#/docs/multilingual>
- Martins, B., Borbinha, J., Pedrosa, G., Gil, J., & Freire, N. (2007). Geographically-aware information retrieval for collections of digitized historical maps. *Proceedings of GIR '07*.
- Milleville, K., Verstockt, S., & Van de Weghe, N. (2022). Automatic georeferencing of topographic raster maps. *International Journal of Geo-Information*, 11(7), 387. <https://doi.org/10.3390/ijgi11070387>
- Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems*, 67, 6969–7055. <https://doi.org/10.1007/s10115-025-02429-y>
- Mollema, W. J. T. (2025). A taxonomy of epistemic injustice in the context of AI and the case for generative hermeneutical erasure. *AI Ethics*, 5, 5535–5555. <https://doi.org/10.1007/s43681-025-00801-w>
- Mõisja, K. (2019). Kaardi kujundamine. J. Roosaare, K. Mõisja, & R. Aunap, *Geoinformaatika. Õpik kõrgkoolidele* (lk 584–613). Tartu Ülikooli kirjastus.
- Naemi, F., Ghods, V., & Khalesi, H. (2022). Scene text detection and recognition: a survey. *Multimedia Tools and Applications*, 81, 20255–20290. <https://doi.org/10.1007/s11042-022-12693-7>
- Närvivõrkude ja masinõppe sõnastik (kuupäev puudub). Vaadatud 8.2.2026 <http://datasci.ee/masinoppe-sonastik/>
- Olson, R., Kim, J., & Chiang, Y. (2024). Automatic search of multiword place names on historical maps. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data (GeoSearch '24)*, 9–12. Association for Computing Machinery.

[OOS] = Eesti Keele Instituudi Onomastika oskussõnastik (2024). Vaadatud 22.10.2025
<https://sonaveeb.ee/search/unif/dlall/oos/nimesiire/1/est>

Orasmaa, S., Käärik, R., Vilo, J., & Hennoste, T. (2010). Information retrieval of word form variants in spoken language corpora using generalized edit distance. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA). <https://doi.org/10.63317/22tyay67w68d>

Pajusalu, K., Hennoste, T., Niit, E., Päll, P., & Viikberg, J. (2018). *Eesti murded ja kohanimed* (3., kohendatud ja täiendatud trükk).

Pall, V. (1977). *Põhja-Tartumaa kohanimed II*. Eesti NSV Teaduste Akadeemia Keele ja kirjanduse instituut. Valgus.

Pall, V. (1969). *Põhja-Tartumaa kohanimed I*. Eesti NSV Teaduste Akadeemia Keele ja kirjanduse instituut. Kirjastus Valgus.

Pertelson, A. (2025). *Viljandimaa asustusnimede tähendusväli ning ruumiline ja vanuseline muster*. [Bakalaureusetöö, Tartu Ülikool].

Poudroux, J., Gonzato, J.-C., Pereira, A., & Guitton, P. (2007). Toponym recognition in scanned color topographic maps. *Proceedings of the 9th international conference on document analysis and recognition*, 531–535. <https://doi.org/10.1109/ICDAR.2007.4378766>

Päll, P. (2008). *Talunimede muutumine*. VI Kohanimepäev Võru, 20. XI 2008. Vaadatud 16.4.2025 <https://arhiiv.eki.ee/knn/knp0811.htm>

Rahvusarhiiv. (2023, 11. oktoober). *Uurijatund asundustaludest* [Video]. YouTube. Vaadatud 18.4.2026 <https://www.youtube.com/watch?v=FzxulifXoi8>

Rahvusarhiiv. (2021, 14. aprill). *Taluajaloo uurimine arhiivis*. Vaadatud 10.10.2025 <https://www.ra.ee/wp-content/uploads/2021/04/Taluajaloo-uurimine-arhiivis-14.04.2021-slaidid-1.pdf>

Rosebrock, A. (2021, 12. mai). *Adaptive Thresholding with OpenCV* (cv2.adaptiveThreshold). Vaadatud 31.3.2026 <https://pyimagesearch.com/2021/05/12/adaptive-thresholding-with-opencv-cv2-adaptivethreshold/>

Saar, E. (2008a). *Võrumaa kohanimede analüüs enamlevinud nimeosade põhjal ja traditsioonilise kogukonna nimesüsteem*. [Doktoritöö, Tartu Ülikool].

Saar, E. (2008b). *Ammõtlikõ talunimmi keelelidsõ ainõ muutuminõ Võromaal. Ametliku talunimistu muutumine keelevahetuse tingimustes*. Kohanimepäev Võrus 20. november 2008. Vaadatud 16.4.2025 <https://arhiiv.eki.ee/knn/knp0811.htm>

Saar, E. (2002). Räpina ja Vahtsõliina kohanimed. Sünkrooniline ülevaade ja andmebaas. M. Faster, E. Saar, & P. Päll (toim), *Võromaa kotussõnimmist* (lk 113–212). Võro instituudi toimõndusõq 13.

Schlegel, I. (2021). Automated extraction of labels from large-scale historical maps. *AGILE: GIScience Series*, 2(12). <https://doi.org/10.5194/agile-giss-2-12-2021>

Sink, G. (2013). *Valjala kihelkonna asustus ja kohanimed ajalooliste kaartide põhjal*. [Bakalaureusetöö, Tartu Ülikool].

- Suurna, R., & Sisask, E. (2010). *GIS ja kartograafia alused*. Riiklik Eksami- ja Kvalifikatsioonikeskus.
- Sügis, E., Tampuu, A., Aljanaki, A., Fišel, M., & Kull, M. (2024/2025). *Praktiline andmeteadus. Kõrgkooliõpik*. Tartu Ülikooli arvutiteaduse instituut.
- Zhou, S. (2026). A survey on transformer for optical character recognition. *Academic Journal of Science and Technology*, 19(2), 395–400. <https://doi.org/10.54097/fdgc6p75>
- Zhu, Y., Yao, C., & Bai, X. (2016). Scene text detection and recognition: recent advances and future trends. *Frontiers of Computer Science*, 10(1), 19–36. <https://doi.org/10.1007/s11704-015-4488-0>
- Tauschek, G. (1929/1935). *Reading machine* (U.S. Patent No. 2026330A). U.S. Patent and Trademark Office. Vaadatud 25.4.2026 <https://patentimages.storage.googleapis.com/ed/c8/ed/4d6d9433b44ee3/US2026330.pdf>
- Tessdoc (kuupäev puudub). Vaadatud 29.3.2026 <https://tesseract-ocr.github.io/tessdoc>
- Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Zie, S. (2024). Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (lk 9568–9578). <https://doi.org/10.1109/CVPR52733.2024.00914>
- Troska, G. (1998/2008). Külad ja külaelu. A. Viires, & E. Vunder (koost ja toim), *Eesti rahvakultuur* (2. väljaanne, lk 199–212). Eesti entsüklopeediakirjastus.
- Troska, G. (1995). *Talunimed läbi aegade*. Teaduste Akadeemia Kirjastus.
- Troska, G., & Šchlögina, N. (1976). Eesti külavormide klassifitseerimisest. I. Sõnajalg (toim), *Etnograafiamuuseumi aastaraamat XXIX* (lk 101–125). Valgus.
- Tõnisson, E., Palts, T., Tõnisson, K., Meier, H., Säde, M., Luberg, A., Lorenz, B., Kivisalu, E., Antoi, M., & Krusberg, S. M. (2019). *Tarkvaraarendus* (2. trükk). Vaadatud 10.2.2026 <https://web.htk.tlu.ee/digitaru/tarkvara2/>
- Tärk, T. (2010). *Kiili valla asustusnimed*. [Magistritöö, Tartu Ülikool].
- Waheed, S., Ferrarini, B., Milford, M., Ramchurn, S. D., & Ehsan, S. (2025). Image-based geo-localization for robotics: are black-box vision-language models there yet?. arXiv. <https://doi.org/10.48550/arXiv.2501.16947>
- Wang, H., Pan, C., Guo, X., Ji, C., & Deng, K. (2025). From object detection to text detection and recognition: a brief evolution history of optical character recognition. *WIREs Computational Statistics*, 13(5). <https://doi.org/10.1002/wics.1547>
- Wang, J. (2023). A study of the OCR development history and directions of development. *Highlights in Science, Engineering and Technology*, 72, 409–415. <https://doi.org/10.54097/bm665j77>
- Wang, X.-F., He, Z.-H., Wang, K., Wang, Y.-F., Zou, L., & Wu, Z.-Z. (2023). A survey of text detection and recognition algorithms based on deep learning technology. *Neurocomputing*, 556. <https://doi.org/10.1016/j.neucom.2023.126702>
- Äriregister (2025). Vaadatud 9.9.2025 <https://ariregister.rik.ee/>

Yang, L., Ergu, D., Cai, Y., Liu, F., & Ma, B. (2022). A review of natural scene text detection methods, *Procedia Computer Science*, 199, 1458–1465.
<https://doi.org/10.1016/j.procs.2022.01.185>

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Zu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, 11(12).
<https://doi.org/10.1093/nsr/nwae403>

Lisad

Lisa 1. Tehisaru kasutamise aruanne

Erinevaid tuvastusmudeleid on otsitud mh tehisaruga modelleerimaks nn tavakasutaja valikut. Selle abil on kirjutatud skripte ning lahendatud tehnilisi probleeme, näiteks otsitud vigu skriptidest keelemudelite väljastatud koordinaatidest. Viimane kasutusala on ka tekstis võimalusena välja toodud. Kõik skriptid on aga autori poolt üle vaadatud. Lisaks tehisaru abil on otsitud kirjandust, kuid kõik viidatud kirjandus on autori poolt läbi loetud ja refereeritud. Töö tekst on tervikuna autori poolt kirjutatud ja autor vastutab võimalike vigade eest.

Kasutatud tehisaru mudelid on olnud OpenAI ChatGPT-5 ning Tartu NLP töörühma tehisaru hindamise rakendus Tehisaru baromeeter¹⁶, milles on 60 erinevat mudelit, millest loositakse välja kaks, mis annavad vastuse.

Lisa 2. Tasuliste mudelite viip

„This map is in Estonian. Read all horizontal text labels from this historical map tile.

Detect even faint or partially visible labels.

Only extract labels written horizontally.

Return JSON exactly like this:

```
[{"name":"Tartu","x":111,"y":222},  
{"name":"Tallinn","x":333,"y":333}]“
```

¹⁶ <https://baromeeter.tartunlp.ai/>

Lisa 3. Kombineeritud meetodi viip

„This map is in Estonian. Grid contains labels arranged row-wise.

Each cell is indexed left-to-right, top-to-bottom:

0 1 2 3 4

5 6 7 8 9

...

Do NOT merge text from multiple cells.

Each cell must be treated independently.

Example:

If a name is split like:

cell 3: "Garmisch-"

cell 4: "Partenkirchen"

Return:

```
[{"cell": 3, "name": "Garmisch-"},  
{"cell": 4, "name": "Partenkirchen"}]
```

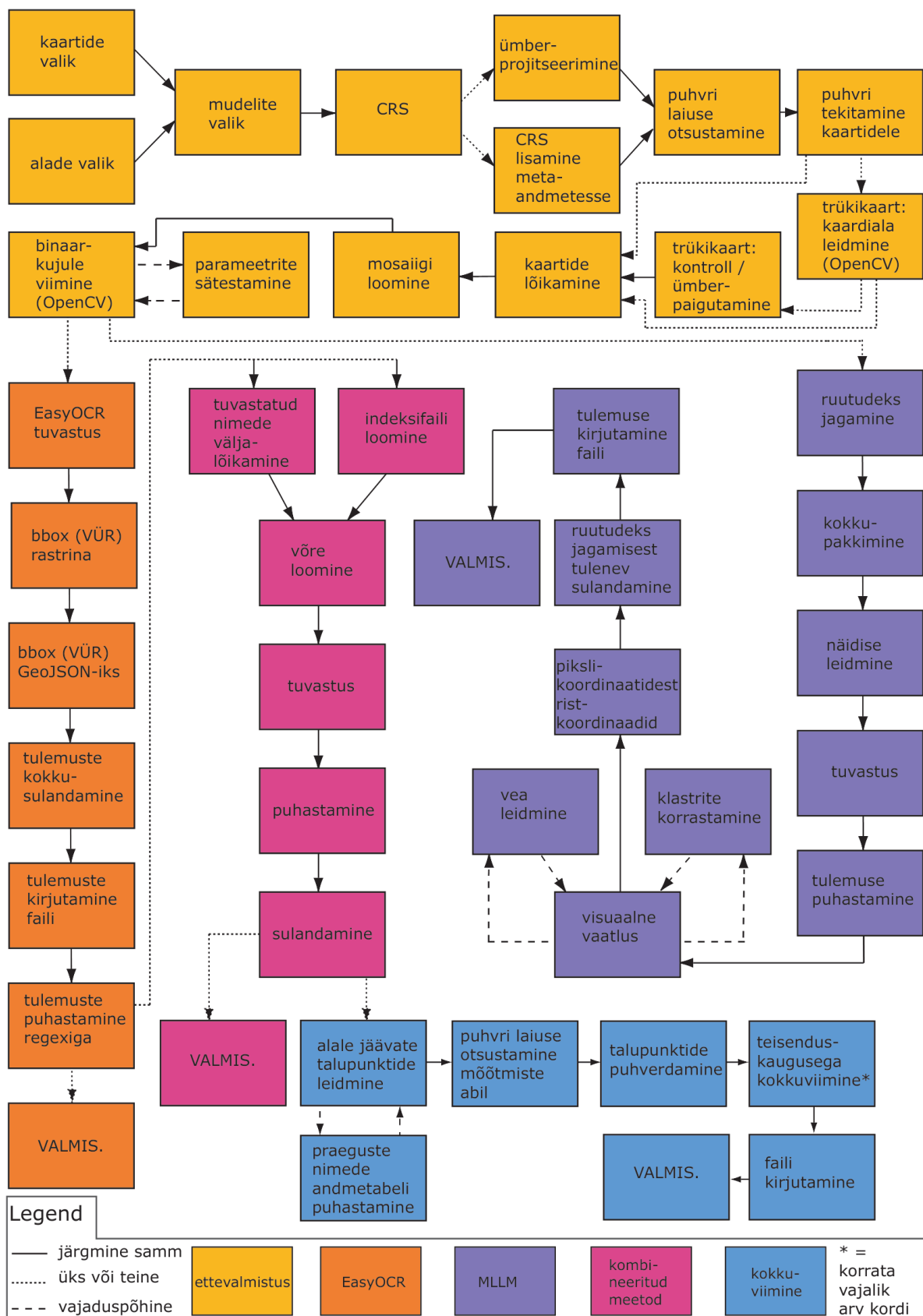
Return JSON exactly like this:

```
[{"cell":0,"name":"..."},  
{"cell":1,"name":"..."}]
```

Lisa 4. Tuvastusskript

Skript on ühe näidiskaardilehe jaoks ja see on leitav Jupyter Notebookina aadressilt <https://github.com/spatsem/maka/>.

Lisa 5. Töövoog



Lisa 6a. Tulemuste F1-skoorid teisenduskaugusega 0

Rasvases kirjas $\geq 0,95$, allajoonitud $\leq 0,25$.

TESTALA	KAART	MUDEL	TALUNIMI TK	VASTEID (mitte NULL)	TÄPSUS (Levenshtein=0, puhastamata)	TÄPSUS (Levenshtein=0, puhastatud)	FALSE POSITIVE	FALSE NEGATIVE	TÄPSUS (PRECISION)	SAAGIS (RECALL)	F1	MÄRKUS
54082	kehtiv EPK	EasyOCR	126	102	47	52	50	74	0,51	0,41	0,46	
54654	kehtiv EPK	EasyOCR	46	41	22	26	15	20	0,63	0,57	0,60	
62471	kehtiv EPK	EasyOCR	83	76	54	56	20	27	0,74	0,67	0,70	
63644	kehtiv EPK	EasyOCR	78	69	41	44	25	34	0,64	0,56	0,60	
64593	kehtiv EPK	EasyOCR	74	65	38	38	27	36	0,58	0,51	0,55	
54082	EPK 20T	EasyOCR	50	20	4	6	14	44	0,30	<u>0,12</u>	0,17	C15
54082	EPK 20T	EasyOCR	50	17	3	5	12	45	0,29	<u>0,10</u>	0,15	C11
54654	EPK 20T	EasyOCR	94	89	62	66	23	28	0,74	0,70	0,72	C15
54654	EPK 20T	EasyOCR	94	89	58	63	26	31	0,71	0,67	0,69	C11
62471	EPK 20T	EasyOCR	46	45	29	35	10	11	0,78	0,76	0,77	
63644	EPK 20T	EasyOCR	43	42	24	28	14	15	0,67	0,65	0,66	
64593	EPK 20T	EasyOCR	78	76	33	41	35	37	0,54	0,53	0,53	
54082	EPK 10T	EasyOCR	111	70	46	53	17	58	0,76	0,48	0,59	
54654	EPK 10T	EasyOCR	84	80	74	75	5	9	0,94	0,89	0,91	
62471	EPK 10T	EasyOCR	69	68	57	63	5	6	0,93	0,91	0,92	
63644	EPK 10T	EasyOCR	58	53	35	43	10	15	0,81	0,74	0,77	
64593	EPK 10T	EasyOCR	44	44	39	39	5	5	0,89	0,89	0,89	
54654	kehtiv EPK	Gemini 2.5	46	37	35	35	2	11	0,95	0,76	0,84	6x6
54082	EPK 20T	Gemini 2.5	50	34	25	25	9	25	0,74	0,50	0,60	C15, 6x6
54082	EPK 10T	Gemini 2.5	111	60	44	44	16	67	0,73	0,40	0,51	6x6
54082	kehtiv EPK	Gemini 3	126	88	84	84	4	42	0,95	0,67	0,79	3x3
54082	EPK 20T	Gemini 3	50	41	36	36	5	14	0,88	0,72	0,79	C15, 3x3
54082	EPK 20T	Gemini 3	50	42	35	35	7	15	0,83	0,70	0,76	C15, 6x6
64593	EPK 20T	Gemini 3	78	35	27	27	8	51	0,77	0,35	0,48	6x6
54082	EPK 20T	Qwen	50	37	26	26	11	24	0,70	0,52	0,60	C15, 6x6
63644	kehtiv EPK	Qwen	78	52	46	46	6	32	0,88	0,59	0,71	6x6
63644	EPK 10T	Qwen	58	39	28	33	6	25	0,85	0,57	0,68	6x6
54082	kehtiv EPK	kombineeritud	126	102	74	74	28	52	0,73	0,59	0,65	
54082	EPK 10T	kombineeritud	111	71	67	67	4	44	0,94	0,60	0,74	
54082	EPK 20T	kombineeritud	50	19	11	11	8	39	0,58	<u>0,22</u>	0,32	C15
54654	kehtiv EPK	kombineeritud	46	41	36	36	5	10	0,88	0,78	0,83	
54654	EPK 10T	kombineeritud	84	80	80	80	0	4	1,00	0,95	0,98	

TESTALA	KAART	MUDEL	TALUNIMI TK	VASTEID (mitte NULL)	TÄPSUS (Levenshtein=0, puhastamata)	TÄPSUS (Levenshtein=0, puhastatud)	FALSE POSITIVE	FALSE NEGATIVE	TÄPSUS (PRECISION)	SAAGIS (RECALL)	F1	MÄRKUS
54654	EPK 20T	kombineeritud	94	89	87	87	2	7	0,98	0,93	0,95	
62471	kehtiv EPK	kombineeritud	83	76	68	68	8	15	0,89	0,82	0,86	
62471	EPK 10T	kombineeritud	69	68	68	68	0	1	1,00	0,99	0,99	
62471	EPK 20T	kombineeritud	46	45	44	44	1	2	0,98	0,96	0,97	
63644	kehtiv EPK	kombineeritud	78	69	58	58	11	20	0,84	0,74	0,79	
63644	EPK 10T	kombineeritud	58	53	48	48	5	10	0,91	0,83	0,86	
63644	EPK 20T	kombineeritud	43	42	40	40	2	3	0,95	0,93	0,94	
64593	kehtiv EPK	kombineeritud	74	65	55	55	10	19	0,85	0,74	0,79	
64593	EPK 10T	kombineeritud	44	44	44	44	0	0	1,00	1,00	1,00	
64593	EPK 20T	kombineeritud	78	76	67	67	9	11	0,88	0,86	0,87	

Lisa 6b. Tulemuste F1-skoorid teisenduskaugustega 0 ja 1
Rasvases kirjas $\geq 0,95$, allajoonitud $\leq 0,25$.

TESTALA	KAART	MUDEL	TALUNIMI TK	VASTEID (mitte NULL)	TÄPSUS (Levenshtein=0, puhastamata)	TÄPSUS (Levenshtein= ≤ 1 , puhastatud)	FALSE POSITIVE	FALSE NEGATIVE	TÄPSUS (PRECISION)	SAAGIS (RECALL)	F1	MÄRKUS
54082	kehtiv EPK	EasyOCR	126	102	47	75	27	51	0,74	0,60	0,66	
54654	kehtiv EPK	EasyOCR	46	41	22	37	4	9	0,90	0,80	0,85	
62471	kehtiv EPK	EasyOCR	83	76	54	70	6	13	0,92	0,84	0,88	
63644	kehtiv EPK	EasyOCR	78	69	41	57	12	21	0,83	0,73	0,78	
64593	kehtiv EPK	EasyOCR	74	65	38	54	11	20	0,83	0,73	0,78	
54082	EPK 20T	EasyOCR	50	20	4	11	9	39	0,55	<u>0,22</u>	0,31	C15
54082	EPK 20T	EasyOCR	50	17	3	9	8	41	0,53	0,18	0,27	C11
54654	EPK 20T	EasyOCR	94	89	62	84	5	10	0,94	0,89	0,92	C15
54654	EPK 20T	EasyOCR	94	89	58	81	8	13	0,91	0,86	0,89	C11
62471	EPK 20T	EasyOCR	46	45	29	45	0	1	1,00	0,98	0,99	
63644	EPK 20T	EasyOCR	43	42	24	40	2	3	0,95	0,93	0,94	
64593	EPK 20T	EasyOCR	78	76	33	62	14	16	0,82	0,79	0,81	
54082	EPK 10T	EasyOCR	111	70	46	63	7	48	0,90	0,57	0,70	
54654	EPK 10T	EasyOCR	84	80	74	78	2	6	0,98	0,93	0,95	
62471	EPK 10T	EasyOCR	69	68	57	68	0	1	1,00	0,99	0,99	
63644	EPK 10T	EasyOCR	58	53	35	47	6	11	0,89	0,81	0,85	
64593	EPK 10T	EasyOCR	44	44	39	43	1	1	0,98	0,98	0,98	Ainus Rõõsimagi.
54654	kehtiv EPK	Gemini 2.5	46	37	35	37	0	9	1,00	0,80	0,89	6x6
54082	EPK 20T	Gemini 2.5	50	34	25	34	0	16	1,00	0,68	0,81	C15, 6x6
54082	EPK 10T	Gemini 2.5	111	60	44	49	11	62	0,82	0,44	0,57	6x6
54082	kehtiv EPK	Gemini 3	126	88	84	87	1	39	0,99	0,69	0,81	3x3
54082	EPK 20T	Gemini 3	50	41	36	41	0	9	1,00	0,82	0,90	C15, 3x3
54082	EPK 20T	Gemini 3	50	42	35	41	1	9	0,98	0,82	0,89	C15, 6x6
64593	EPK 20T	Gemini 3	78	35	27	29	6	49	0,83	0,37	0,51	6x6
54082	EPK 20T	Qwen	50	37	26	33	4	17	0,89	0,66	0,76	C15, 6x6

TESTALA	KAART	MUDEL	TALUNIMI TK	VASTEID (mitte NULL)	TÄPSUS (Levenshtein=0, puhastamata)	TÄPSUS (Levenshtein= ≤ 1 , puhastatud)	FALSE POSITIVE	FALSE NEGATIVE	TÄPSUS (PRECISION)	SAAGIS (RECALL)	F1	MÄRKUS
63644	kehtiv EPK	Qwen	78	52	46	50	2	28	0,96	0,64	0,77	6x6
63644	EPK 10T	Qwen	58	39	28	35	4	23	0,90	0,60	0,72	6x6
54082	kehtiv EPK	kombineeritud	126	102	74	82	20	44	0,80	0,65	0,72	
54082	EPK 10T	kombineeritud	111	71	67	68	3	43	0,96	0,61	0,75	
54082	EPK 20T	kombineeritud	50	19	11	14	5	36	0,74	0,28	0,41	C15
54654	kehtiv EPK	kombineeritud	46	41	36	40	1	6	0,98	0,87	0,92	
54654	EPK 10T	kombineeritud	84	80	80	80	0	4	1,00	0,95	0,98	
54654	EPK 20T	kombineeritud	94	89	87	87	2	7	0,98	0,93	0,95	
62471	kehtiv EPK	kombineeritud	83	76	68	72	4	11	0,95	0,87	0,91	
62471	EPK 10T	kombineeritud	69	68	68	68	0	1	1,00	0,99	0,99	
62471	EPK 20T	kombineeritud	46	45	44	45	0	1	1,00	0,98	0,99	
63644	kehtiv EPK	kombineeritud	78	69	58	63	6	15	0,91	0,81	0,86	
63644	EPK 10T	kombineeritud	58	53	48	50	3	8	0,94	0,86	0,90	
63644	EPK 20T	kombineeritud	43	42	40	40	2	3	0,95	0,93	0,94	
64593	kehtiv EPK	kombineeritud	74	65	55	59	6	15	0,91	0,80	0,85	
64593	EPK 10T	kombineeritud	44	44	44	44	0	0	1,00	1,00	1,00	
64593	EPK 20T	kombineeritud	78	76	67	70	6	8	0,92	0,90	0,91	

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Maria Anttila,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

Talunimede tuvastamise automatiseerimine Eesti põhikaartide näitel,

mille juhendajad on Raivo Aunap, Siim Orasmaa ja Ardi Tampuu,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Maria Anttila

25.05.2026