TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Karoliina Jõgi

NIMI-, OMADUS- JA MÄÄRSÕNAFRAASILISED PÜSIÜHENDID EESTIKEELSETES TEKSTIDES

Bakalaureusetöö

Juhendaja PhD Kadri Muischnek

TARTU 2025

Autorsuse kinnitus

Kinnitan, et olen käesoleva lõputöö ise kirjutanud ning toonud korrektselt välja teiste autorite panuse. Töö on kirjutatud, lähtudes Tartu Ülikooli eesti ja üldkeeleteaduse instituudi lõputöö nõuetest, ning on kooskõlas heade akadeemiliste tavadega.

Karoliina Jõgi

Lühikokkuvõte

Bakalaureusetöö eesmärk oli teada saada esiteks, millised püsiühendid on eri liiki tekstides sagedasemad, peamiselt sõnaliigi ja kokku või lahku kirjutamise järgi, ja teiseks, kuidas on rahvusvahelise UniDive'i intistiatiivi püsiühendite testid kohandatavad eesti keelele. Selleks märgendasin eesti keele Universal Dependecies puudepangas (Zeman jt 2024) iga valitud tekstiliigi kohta (ajakirjanduslikud, ilukirjanduslikud ja populaarteaduslikud tekstid) viiesajas lauses nimi-, omadus- ja määrsõnafraasilisi püsiühendeid ning nende alaliike.

Töö tulemusteks on eesti keele jaoks kohandatud püsiühendite testid, nende testide alusel märgendatud korpus ning teadmised püsiühendite esinemissagedusest, sh püsiühendiliikide ja tekstiliikide kaupa. Märgendamiseks kohandasin UniDive'i keelteüleseid püsiühenditeste (Savary jt *s. a.* (b): 3–8) ja nende järjekorda eesti keele jaoks. Nimisõnafraasilisi püsiühendeid leidis kõige rohkem täiendi lisamise test, modifitseerijafraasilisi aga sõna asendamise test. Kaks kasutatud testidest ei tuvastanud ühtegi nimisõnafraasilist ja kaks testidest ühtegi omadussõnafraasilist püsiühendit.

Kõige rohkem esines nimisõnafraasilisi püsiühendeid, siis määrsõnafraasilisi, deverbaalseid nimisõnalisi, omadussõnafraasilisi, deverbaalseid modifitseerijafraasilisi ja veidi ka asesõnalisi. 70% püsiühendilemmadest esines vaid korra ja üle 80% esines ainult ühes failis üheksast, seega oli püsiühendite esinemissageduste jaotus väga ebaühtlane. Pea 90% leitud püsiühenditest olid liitsõnad ja esinenud liitsõnadest pea kolmandik olid püsiühendid, kuigi unikaalseid lemmasid vaadeldes langes osakaal 14%-le. Tekstiliigiti esines püsiühendeid rohkem populaarteaduslikes tekstides ja vähem ilukirjanduslikes tekstides, aga unikaalseid lemmasid oli enim ajakirjanduslikes tekstides.

Võtmesõnad: püsiühend, nimisõnafraas, omadussõnafraas, määrsõnafraas, korpuse märgendamine

Sisukord

Sissejuhatus 5

1. Püsiühendilisuse tunnused ja testimine 7

2. Püsiühendite käsitlusi erinevatest vaatenurkadest 12

2.1. Keeleteaduslikud käsitlused Eestis 12

2.2. Arvutilingvistika käsitlused 13

3. Püsiühendite märgendamine korpustes 17

3.1. UniDive'i eellased 17

3.2. UniDive'i initsiatiiv ja motivatsioon 17

3.3. Püsiühendid UniDive'i initsiatiivi raamistikus 18

4. Püsiühendite liigituse ja testimise katsetamine eesti keele materjalil 21

4.1. Nimisõnafraasiliste püsiühendite testid 21

4.2. Modifitseerijafraasiliste püsiühendite testid 25

5. Materjal ja meetod 28

5.1. Märgendamisel esile kerkinud probleemkohad 30

6. Tulemuste analüüs 33

6.1. Püsiühendite sagedused ja liigid 34

6.2. Püsiühendid eri tekstiliikides 36

6.3. Püsiühendite testid 38

6.4. Liitsõnalisus 39

Kokkuvõte 41

Kirjandus 44

Nominal, Adjectival and Adverbial Multiword Expressions in Estonian Texts. Summary 47

Lisa 1. Ekraanipilt ühest märgendatud tekstist INCEpTIONis 49

Lisa 2. Ekraanipilt ühest märgendatud tekstist tsv-faili kujul 50

Lisa 3. Tabel vähemalt kaks korda esinenud püsiühenditega. 51

Sissejuhatus

Igas keeles on sõnade ühendeid, mis koos tähendavad midagi muud kui eraldiseisvatena. Püsiühendid ongi sellised püsivad sõnade ühendid, mis kalduvad keele üldistest reeglitest millegi, näiteks semantika või morfoloogia, poolest kõrvale. Näiteks liitsõnas *vikerkaar* on esiosa *viker*, mis esineb väga vähestes liitsõnades ja iseseisva sõnana ei esinegi; ühendi *nii-öelda* struktuur on adverb-verb, aga ühend tervikuna käitub süntaktiliselt adverbina. Kuigi püsiühendeid kasutatakse keeles palju, on nende mõiste ja seega hulga kindlapiiriline määratlemine keeruline, sest oluline osa nende eripärast on ühendi tähenduse erinemine selle komponentide tähendustest, mis võib olla subjektiivne.

Eriti keeruline on nende määratlemine keelteüleselt, mis võimaldaks keeli selles osas võrrelda. Rahvusvahelise UniDive'i initsiatiivi[1] eesmärk on kooskõlastada keeleline mitmekesisus keeletehnoloogia kiire arenguga ja sealhulgas märgendada püsiühendeid keelteüleselt ühtlasena (UniDive *s. a.*). Seetõttu otsin püsiühendeid eri liiki eestikeelsetest tekstidest UniDive'i initsiatiivi raamistiku järgi, et leida, mida sellisele rahvusvahelisele raamistikule toetudes eesti keele püsiühendite kohta öelda saab.

Selle uurimistöö üks eesmärk on välja selgitada, millised püsiühendid on eri liiki tekstides sagedasemad. Peamiselt uuritakse eri liiki (nimi-, omadus- ja määrsõnafraasilised) ja eri kujul (liitsõnaliste, mida selles töös käsitlen sõnaühenditena, või eraldi sõnade kujul) püsiühendite sagedust. Teine eesmärk on teada saada, kas ja kuidas on UniDive'i püsiühendite määramise testid kohandatavad eesti keele püsiühendite tuvastamiseks. Eesmärgi täitmiseks olen püstitanud järgmised uurimisküsimused:

1. Millised püsiühendid esinevad tekstis kõige rohkem, sealhulgas sõnaliigi poolest?
2. Kuidas erineb püsiühendite kasutus erinevates registrites (ajalehetekstid, ilukirjandus- ja populaarteaduslikud tekstid)?
3. Kuidas saab UniDive'i initsiatiivi püsiühenditeste eesti keele jaoks kohandada ja millised testid tuvastavad kõige rohkem püsiühendeid?
4. Kui paljud leitavatest püsiühenditest on liitsõnad ja kui paljud esinevatest liitsõnadest on püsiühendid?

Töö esimeses osas kirjeldan püsiühendite tunnuseid ning teises osas Eesti autorite ja arvutilingvistika käsitlusi püsiühenditest. Kolmandas peatükis käsitlen püsiühendite märgendamist korpustes. Täpsemalt kirjutan lühidalt kahest UniDive'i eellasest ja UniDive'i

[1] https://unidive.lisn.upsaclay.fr/.

initsiatiivist ning seejärel püsiühendite käsitlusest selles initsiatiivis. Neljandas osas kirjutan püsiühendite tuvastamiseks kasutatud testidest ja nende kohandamisest. Viiendas osas kirjeldan materjali ja meetodit. Lisaks toon välja märgendamisel esile kerkinud probleemkohad. Kuuendas peatükis analüüsin tulemusi neljast uurimisküsimusest lähtuvalt.

1. Püsiühendilisuse tunnused ja testimine

Kuigi püsiühendi võib-olla lihtsaim määratlus on „püsiv sõnade ühend“, ei ole piir püsiühendite ja vabade sõnaühendite vahel selgepiiriline ning ka nende alaliikide kohta käivate terminite kasutus kõigub autoriti nii eesti kui ka inglise keeles. Nii on eesti keeles kasutatud ka näiteks termineid fraseologism, kollokatsioon ja idioom, mida on käsitletud ka püsiühendite alaliikide nimetustena (Õim & Õim 2019: 5, 46), ning inglise keeles näiteks termineid *fixed expressions, idioms, collocations* ja *multiword expressions* (Moon 1998: 2–5; Ramisch & Villavicencio 2022: 652). Lisaks terminitele varieerub ka see, milliste tunnustega püsiühendeid defineeritakse ja kuidas mõne ühendi püsiühendilisust testitakse (Ramisch & Villavicencio 2022: 652; Õim & Õim 2019: 31).

Erinevad autorid on toonud esile erinevaid püsiühendite tunnuseid ning rõhutanud eri tunnuste olulisust. Osad tunnused tulevad eri autorite tekstides rohkem esile kui teised, näiteks enamus mainivad (mitte-)kompositsioonilisust ehk terviku tähenduse tuletatavust selle komponentide tähendustest, näiteks liitsõna *eesmärk* tähendus pole tuletatav sõnade *ees* ja *märk* tähendustest, ning püsivust ja piiratud paindlikkust, mis käivad tihti koos, sest püsivus viitab piiratud paindlikkusele, näiteks *eeskätt* ’eriliselt (just)’ tähendab midagi muud kui *eeskäsi* ’kaardimängus alustaja’. Samuti mainitakse tihti vähemalt kahe leksikaalse komponendi vajalikkust ja sagedat kasutust. (nt Moon 1998: 7–8; Gries 2008: 4–6; Ramisch & Villavicencio 2022: 652–654) Ilmselt ongi üks püsiühendite mõiste varieerumise põhjusi see, et eri uurijad omistavad sellele erinevaid tunnuseid ja piiritlevad seda veidi erinevalt, mis tuleneb ka erinevatest uurimishuvidest.

Stefan Th. Gries (2008: 4–6) pakub välja kuus kriteeriumit, mida uurijad võiksid oma püsiühendi definitsioonides täpsustada:

1) mida püsiühendi elementideks peetakse, näiteks lemma või grammatiline element;
2) mitu elementi võib või peab püsiühendis olema – selleks on enamasti kaks või rohkem, aga võib olla ka näiteks täpselt kaks;
3) mitu korda peab väljend esinema, et seda püsiühendiks saaks lugeda, kusjuures vähesed uurijad määratlevad seda konkreetse arvuna ja enamik nii, et ühendi esinemissagedus peab ületama juhuse põhjal eeldatava esinemiste hulga, mida arvutatakse enamasti ühendi eri elementide korpuses esinemise sageduse järgi;
4) kui suur on lubatud kaugus elementide vahel ehk kas loevad ainult kõrvuti asetsevad elemendid, mida tehakse enamasti n-grammide ehk *n* hulga kõrvuti asetsevade sõnade

puhul, või võib nende vahel ka teisi sõnu esineda, mis on eriti oluline tegusõnaliste püsiühendite puhul, kus üks osa võib olla päris lause lõpus;

5) kui suur võib olla elementide leksikaalne ja süntaktiline paindlikkus, näiteks kas loevad ainult ühendid, mis on täiesti muutumatud, nagu *käe-jala juures*, või ka sellised, mis võivad teatud määral varieeruda, näiteks ingliskeelne *kick the bucket* (’surema’, sõnasõnalt ’pange lööma’) võib esineda eri aegades, aga mitte passiivis;
6) kui olulised on püsiühendi tähenduslik terviklikkus ja semantiline mitte-kompositsioonilisus ehk terviktähenduse tuletatavus komponentide tähendustest.

Need kriteeriumid lasevad vajaduse järgi definitsioone teha väga kitsaks või laiaks, aga mõned määratlused on rohkem levinud kui teised, näiteks paljudes definitsioonides on mainitud esimest kahte kriteeriumit ehk toodud välja, et püsiühend peab koosnema vähemalt kahest püsivast sõnast (Moon 1998: 8) või lemmast (Savary jt *s. a.* (a): 2; Ramisch & Villavicencio 2022: 652; Ramisch 2023: 9–10). Seega ei sobi grammatilistest kategooriatest koosnevad ühendid, näiteks elatiivis nimisõna ja omadussõna ühendid, nagu *unest segane* ja *vihast punane*.

Ka kolmandat, esinemissageduse kriteeriumit mainitakse tihti ühe püsiühendi tunnusena, küll veidi varieeruva nimetusega, näiteks institutsionaliseerumine (Moon 1998: 7; Granger & Meunire 2008: XIX), konventsionaalsus (Ramisch & Villavicencio 2022: 653; Ramisch 2023: 11) ja usuaalsus (Õim & Õim 2019: 48). Mitte-kompositsioonilised püsiühendid kõlavad loomulikuna just siis, kui neid kuuleb sagedasti (Ramisch & Villavicencio 2022: 653), samas võivad sagedused keele eri allkeeltes ja valdkondades oluliselt varieeruda, eriti kuna neis võivad olla valdkonnaspetsiifilised püsiühendid (Moon 1998: 7). Lisaks on osad püsiühendid konventsionaliseerunud kaua aega tagasi ja kuigi tänapäeval ei pruugi need enam nii sagedased olla, on need siiski kinnistunud (Moon 1998: 7). Näiteks ühendit *eite-taati*, mida saab tõlgendada ka nominatiivina, kuigi vormilt on see partitiivis, enam igapäevaselt ei kasutata, aga on muinasjuttude mõjul veel püsinud. Usuaalsust on tõlgendatud ka kui kasutuses püsivust ehk sotsiaalset püsivust, ilma milleta on teised püsivuse aspektid tähendusetud (Õim & Õim 2019: 48).

Griesi viiendat kriteeriumi käsitlevad mõned püsivuse ja teised paindlikkuse seisukohalt, samuti on kaasatud piiratud varieeruvust morfoloogilises paradigmas, näiteks aspektis, kõneviisis või tegumoes verbide puhul, näiteks väljend *võta näpust* esineb ainult käskivas kõneviisis, ja arvus või käändes noomenite puhul, näiteks *kunstielu* ei saa esineda mitmuses. Piirangud võivad olla ka süntaktilised, leksikaalsed ja semantilised (Õim & Õim 2019: 45), näiteks püsiühendi *maa ja ilm* sõnajärge ei saa muuta ja püsiühendis *kaubamaja* ei saa

komponente asendada semantiliselt sarnaste sõnadega *toode* või *hoone*. Korpusandmed toovad piiranguid esile, samas tõestavad ka teatud varieerumise tavapärasust (Moon 1998: 7), näiteks veidi erinevalt sõnastatud ühendid *omal käel* ja *oma käe peal* on täiesti samatähenduslikud (Õim & Õim 2019: 45). Üks keerukusi püsiühendite tuvastamisel ongi see, et nende püsivus võib palju varieeruda: osad on täiesti jäigad ega muutu üldse ning osad võivad esineda üsna erineval kujul (Ramisch & Villavicencio 2022: 654; Õim & Õim 2019: 53). Püsiühenditel võib erineda ka see, kui paljud ja millised sõnad võivad varieeruda ning kui paljud ja millised sõnad neid asendada võivad, näiteks ühendi *hing läheb välja* puhul võivad varieeruda nii nimisõna, näiteks *hinge* asemel *toss*, kui ka tegusõna, näiteks *läheb välja* asemel *lendab välja* (Õim & Õim 2019: 53).

Samuti on öeldud, et piiratud varieeruvust kutsub esile mitte-kompositsioonilisus (Ramisch & Villavicencio 2022: 653), mis on osa Griesi kuuendast kriteeriumist koos tähendusliku terviklikkusega. Kui kompositsioonilisus on võime tuletada keele tavaliste tuletusreeglite põhjal sõnade kombinatsiooni semantilisi, süntaktilisi ja muid karakteristikuid (Ramisch & Villavicencio 2022: 653), siis mitte-kompositsioonilisus tähendab, et püsiühendeid ei saa tuletada keele tavaliste tuletusreeglite järgi sõnadest või lihtsamatest fraasidest, vaid neid hoiustatakse ja kasutatakse tervikutena (Mel'čuk 2001: 24, 27).

Mitte-kompositsioonilisus on tihedalt seotud idiosünkraatilisusega ehk eripärasusega, mis tähendab keele tavalistest reeglitest erinemist. Tihti peetakse seda semantiliseks kriteeriumiks, aga see võib hõlmata ka grammatilist, leksikaalset, pragmaatilist, süntaktilist, morfoloogilist ja vahetevahel ka statistilist idiosünkraatilisust (Ramisch & Villavicencio 2022: 652). Näiteks *risti ja põiki* on süntaktiliselt omapärane, sest selle asemel ei saa öelda *põiki ja risti*, *näojooned* morfoloogiliselt, sest ei saa öelda *näojoon*, ja *vikerkaar* leksikaalselt, sest *viker* esineb väga vähestes sõnades ja iseseisva sõnana ei esinegi. Lisaks varieeruvad arvamused sellest, kas ainult statistiliselt ehk koosesinemise sageduse põhjal eripärased ühendid, mida kutsutakse tihti kollokatsioonideks, loevad püsiühenditena või mitte (Ramisch 2023: 10). Idiosünkraatilisus on üks võimalikest püsiühendilisuse testimise viisidest. Näiteks leksikaalset ja semantilist idiosünkraatilisust saab testida, proovides mõnda püsiühendi koosseisus olevat sõna asendada mõne seotud sõnaga, nagu sünonüüm, antonüüm, hüperonüüm või hüponüüm. Morfoloogilist ja süntaktilist idiosünkraatilisust saab testida, proovides kasutada sama käitumist süntaktiliselt sarnastes konstruktsioonides, näiteks saab inglise keeles öelda *by and large* 'üldiselt' (sõnasõnalt 'kõrval ja suur'), aga mitte *by and short* (sõnasõnalt 'kõrval ja väike'). (Ramisch & Villavicencio 2022: 654) Eestikeelseks näiteks sobib ehk *eelkõige* ja selle eeskujul olematu liitsõna *järelkõige.*

Samas on mitte-kompositsioonilisuse kriteeriumit kritiseerinud Rosamund Moon (1998: 8) hindajast ja ajahetkest sõltumise tõttu ning Maria Helena Svensson (2008: 81) selle ebaselguse ja -täpsuse tõttu. Ebatäpsust tekitab asjaolu, et mitte-kompositsioonilisust on võimalik tõlgendada eri vaatenurkadest ja tihti ei täpsustata, millist või milliseid vaatenurki kasutatakse. Svensson (2008: 82–86) toob välja neli eristust:

1) motiveeritavus ehk kui lugejale on juba teada ühendi tähendus, siis kas tal on võimalik leida, et see on ühendi osade tähendustest motiveeritud;
2) läbipaistvus ehk kui lugeja näeb ühendit esimest korda ja teab kõiki selles sisalduvaid sõnu, siis kas ta mõistab selle ühendi tähendust;
3) analüüsitavus ehk kas ühendi iga sõna panustab selle ühendi terviklikku tähendusse;
4) otsesus/sõnasõnalisus ehk kas ühendi otsene tõlgendus oleks püsiühendi tähenduse loogiline vastandus, absurdsus või tähendusetu.

Ta toob ka välja, et läbipaistvus ja motiveeritus sõltuvad lugejast, ning väidab, et ükski neist pole ei vajalik ega ka piisav kriteerium püsiühendi defineerimiseks, eriti kuna ühendid võivad olla ka osaliselt kompositsioonilised. Lisaks võib ühendil olla entsüklopeediline kompositsioonilisus, mille puhul on see nende nelja kriteeriumi järgi kompositsiooniline, aga sellel on ka lisatähendus, näiteks *the White House* on tegelikult ka valge maja, aga selle all mõeldakse ka USA valitsust või poliitikat. Eestikeelseks näiteks sobib ehk *raudtee*, sest see on küll tehtud rauast, aga põhiline tähendus on see, et selle peal sõidavad rongid. Samuti toob ta välja selle, et tihti ei täpsustata, mida sõna tavalise tähenduse all silmas peetakse, näiteks kas vanimat või sagedasimat või otsest tähendust, eriti kuna otsene tähendus pole alati prototüüpne tähendus. (Svensson 2008: 88–90)

Ma nõustun, et mitte-kompositsioonilisus ja sõna tavaline tähendus võivad kohati olla ebaselged ning paljud loevadki mitte-kompositsioonilisust skalaarseks tunnuseks (nt Ramisch & Villavicencio 2022: 652–653; Õim & Õim 2019: 46; Moon 1998: 23). Lisaks keskendub Svenssoni analüüs semantilisele mitte-kompositsioonilisusele, mille kõrval arvestatakse tihti ka teist laadi mitte-kompositsioonilisust, nagu morfoloogilist või süntaktilist (Moon 1998: 8; Ramisch & Villavicencio 2022: 652). Moon (1998: 8) soosib tõlgendust, et komponentidel võivad ühendites olla tavatähendusest erinevad tähendused, seega võib neile vahel leida selgitusi, analoogiaid ja kasutusi teistes ühendites.

Peale mainitud tunnuste on ka teisi, mis pole nii levinud. Näiteks fonoloogiline kriteerium, mille järgi võib intonatsioon aidata kontekstis eristada ühendi otsest ja kaudset tähendust, viimase puhul võib kestus lühem olla, mis kinnitab ka ühendi terviklikkust (Moon 1998: 9). Eesti keeles illustreerivad seda hästi osad liitsõnad, näiteks *suurlinn* või *lasteaed* on veidi teise

hääldusega kui *suur linn* või *laste aed*. Teine vähemlevinud tunnus on tõlkimatus ehk püsiühendi sõnasõnaline tõlkimise ebaloomulik tulemus võib aidata kinnitada, et tegu on püsiühendiga, aga vahel ei saa ka regulaarseid, mitte-püsiühendilisi kombinatsioone otse tõlkida keelte struktuuride erinevuse tõttu ning osad püsiühendid on mitmes keeles olemas (Ramisch & Villavicencio 2022: 654), näiteks *must auk* ja *black hole* ning *igaüks* ja *everyone*.

Eelneva põhjal võib öelda, eri uurijad defineerivad ja tähtsustavad püsiühendite eri tunnuseid veidi erinevalt, tihti uurimisvaldkondadest lähtuvalt. Lõpuks ei ole ühte või mitut kindlat kriteeriumit, mis teeks ühendi püsiühendiks, vaid on tunnused, mis on eri püsiühendites erineval määral esil (Moon 1998: 9).

2. Püsiühendite käsitlusi erinevatest vaatenurkadest

2.1. Keeleteaduslikud käsitlused Eestis

Eesti keeles on püsiühendit defineeritud kui „püsiv tavapärane sõnade ühend keeles“ (Erelt, Erelt & Ross 2020: 608) ja „kahe või enama sõna(vormi) ühend, mida mingi tähenduse väljendamiseks on tavaks koos kasutada“ (Muischnek & Kaalep 2009: 157). Mõlemad definitsioonid on üsna laiad ja näiteks ühendite süntaktilise struktuuri, samal kujul püsivuse ja tähenduse moodustumise viisi varieeruvuse tõttu on püsiühendeid samuti keeruline üheselt määratleda (Muischnek & Kaalep 2009: 158).

Eesti keele käsiraamatus (Erelt, Erelt & Ross 2020: 608) on peatükk fraseoloogia kohta, kus defineeritakse püsiühendi ehk fraseemi mõistet kui „püsiv tavapärane sõnade ühend keeles“ ja lähedast fraseologismi mõistet kui „keeles laialt käibiv püsiv tavapärane sõnade ühend, millele on omane osade tähenduslik kokkukuulumine ning hrl ka metafoorsus“, kusjuures fraseologism on tihti stiililiselt markeeritud. Neil mõistetel võib olla seega raske vahet teha. Üks viis, kuidas neid eristada, on ehk fraseologismi metafoorsus, näiteks *kits kahe heinakuhja vahel* ja *midagi mäekõrguselt ületama* on pigem fraseologismid kui püsiühendid. Fraseologismide sekka kuuluvaid idioome ja võrdluseid võib samuti püsiühenditeks pidada, kuigi kõnekäänud ja vanasõnad on enamasti püsiühendina määratlemiseks liiga pikad, olles tihti terved laused, näiteks *hommik on õhtust targem*. Samas mainitakse „Eesti keele käsiraamatus“ (Erelt, Erelt & Ross 2020: 42), et leksikoloogia uurib muu hulgas sõnade „püsiühendeid ehk fraseologisme“, nii et need on tõesti väga lähedased mõisted.

Kuna püsiühendite terminite kasutus on nii varieeruv, siis täpsustan, et kasutan selles töös mõistet püsiühend, mitte fraseologism, sest seda on keeleteaduses kasutatud ingliskeelse termini *multiword expression* tõlkena (Muischnek & Kaalep 2009: 157) ja vastavat ingliskeelset mõistet kasutatakse töös kasutatavates püsiühendite tuvastamise testides. Fraseologismi terminit kasutatakse peamiselt folkloristikas.

Eestis on uuritud näiteks võrdluste struktuuripõhiseid liike, fraseologismide etümoloogiat, morfoloogiat, süntaksit ja semantikat, erinevate allikate fraseologisme (Õim & Õim 2019: 24–27). Fraseoloogia tänapäeva uurimisprobleemid on nende klassifitseerimine erinevatel alustel, millega on juba kaua tegeletud, nende vormiline varieerumine, semantiline eripära ja selle leksikograafiline kirjeldamine. Samuti uuritakse nende süntaktilist ja laiemalt diskursiivset käitumist, mis on tihedalt seotud ka fraseologismide psühholingvistilise aspekti ja autorifraseoloogiaga. Lisaks on üheks uuemaks uurimisprobleemiks fraseologismide

automaatne leidmine tekstist. (Õim & Õim 2019: 17, 23, 27) Kuigi uuritakse pigem fraseologisme, saab nende abil ka veidi püsiühendite kohta teada.

Üliõpilastöödena on eesti keele püsiühendeid uuritud pigem verbikeskselt (vt nt Aedmaa 2019). Lisaks on uuritud püsiühenditele sarnaseid üksusi fraseologisme, näiteks nende tundmise vaatenurgast (vt nt Õunap 2010; Mensalo 2024) ja tõlkimise vaatenurgast (vt nt Dovgan 2013; Metsla 2014).

2.2. Arvutilingvistika käsitlused

Arvutilingvistika seisukohalt on püsiühendite käsitlemine keeruline, aga vajalik, eriti kuna need moodustavad suure osa keelest (Muischnek & Kaalep 2009: 158; Moon 1998: 64–68) ning, nagu mainitud 1. peatükis, nende süntaktiline ja morfoloogiline paindlikkus ning tähenduse kompositsioonilisus varieerub tohutult. See teeb raskeks lihtsa leksikonipõhise lähenemise, kus tehakse lihtsalt loend püsiühenditest ja nende tähendustest. Näiteks saab öelda nii *omal käel* kui ka *oma käe peal* 'iseseisvalt', aga ainult *tuule peal* 'ebakindlas olukorras', mitte *tuulel*, kuigi neid püsiühendi variante eristab ainult käändelõpu või kaassõna kasutamine. Püsiühendeid tekstist tuvastades või neid sisaldavat teksti genereerides võivad tekkida muu hulgas järgmised probleemid:

1) üldistuse probleem, näiteks võib genereerimissüsteem arvata, et kuna on olemas väljendid *ajas marru* ja *ajas vihale*, siis võiks olla ka *ajas vimmale*;
2) mitte-kompositsioonilisuse probleem, näiteks kuidas ennustada, et liitsõna *eesmärk* tähendusel pole mingit seost sõnade *ees* ja *märk* tähendustega;
3) paindlikkuse probleem, näiteks kuidas arvestada püsiühendite varieeruvate sõnamuutmisealaste piirangutega, sest ei saa öelda *kunstielud* ega *võis-olla*;
4) diakroonilise varieeruvuse probleem, sest kui püsiühendeid lihtsalt loetleda, siis ei pruugi tehnoloogia toime tulla nende varieerumise ja arenemisega. (Sag jt 2002: 2–3)

Kui jätta püsiühendid tekstis tuvastamata, võivad tekkida probleemid semantilise ja süntaktilise analüüsi kvaliteedi juures, näiteks võib tavaanalüüs tuvastada sõna *aru* lauses „Peeter ei saanud ülesandest aru.“ samaväärse objektina kui sõna *piima* lauses „Peeter ei saanud poest piima“, kuigi see on tegelikult osa väljendverbist *aru saama* (Muischnek & Kaalep 2009: 158). Nende ja muude erinevate probleemide lahendamiseks on leitud mitmeid mehhanisme.

Arvutilingvistika vaatenurgast on kasulik eri tuvastamismehhanismide kasutamiseks püsiühendeid liigitada järgmiste tunnuste järgi:

1) nende püsivuse astme ehk elementide järjekorra püsivuse, kõrvuti asetsemise ja mõne sõna asendatavuse põhjal, näiteks sõna *marru* püsiühendis *marru ajama* saab asendada sõnadega *raevu* või *vihale* ilma tähenduse muutumiseta, aga püsiühend *läbi ja lõhki* esineb ainult selles järjekorras ja nende sõnadega;
2) süntaktilise struktuuri põhjal, näiteks kas see käitub lauses nimi- või omadussõnafraasina;
3) tähenduse moodustumise viisi põhjal, jagades need esiteks mitte-kompositsioonilisteks ühenditeks ehk idioomideks, mida saab veel liigitada näiteks tähenduse läbipaistvuse järgi ehk kas seda esimest korda nähes võib tähendust aimata (läbipaistvus on küll skalaarne ja inimeseti erinev), ja teiseks kompositsioonilisteks ühenditeks ehk kollokatsioonideks. (Muischnek & Kaalep 2009: 158–159)

Püsiühendite arvutipõhisel töötlemisel on kaks peamist etappi: (uute) püsiühendite tuvastamine ja nende märgendamine tekstis. Neid etappe on vaja eraldada näiteks seetõttu, et kui tuvastamisel otsitakse esialgu tihti koos esinevaid sõnu, siis näiteks püsiühendi *järgi vaatama* komponendid võivad koos esineda ka püsiühendit moodustamata, nagu lauses „Statistika järgi vaatab ETV saateid üle 60% eestimaalastest.“ (Muischnek & Kaalep 2009: 160–161) Nende kahe etapi vahele jääb tuvastatud püsiühenditest leksikoni koostamine, mida enamasti kontrollib inimene üle ja mida võib seostada rohkem esimese (Ramisch 2023: 22) või teise etapiga (Muischnek & Kaalep 2009: 160).

Tuvastamise etapis leitakse esmalt ühendikandidaadid ehk üksteise lähedal (nt samas osalauses) esinevad või nt süntaktiliselt seotud sõnaühendid, teiseks leitakse tõenäolised püsiühendid nt sageduse või mõne muu statistilise meetodi põhjal, kolmandaks vaatab inimene tõenäolised püsiühendid üle. Esimeses punktis on mõistlik läheneda püsiühendite eri liikidele erinevalt, näiteks verbide puhul tuleb arvestada, et vaba sõnajärje tõttu võib see esineda nii kujul *üles leidma* kui ka *leidma üles* ning osad verbi laiendid võivad oma käändes ja arvus muutuda, näiteks võib olla nii *pidas kõne* kui ka *ei pidanud kõnet*, samas kui nimisõnafraasilised püsiühendid on alati samas järjekorras, aga nende käändumise võimelisus võib fraaside süntaktilisest struktuurist lähtuvalt varieeruda, näiteks ühendi *hullu lehma tõbi* puhul käändub vaid viimane sõna, *must auk* puhul aga mõlemad sõnad. (Muischnek & Kaalep 2009: 161–163)

Teises punktis saab esiteks kasutada näiteks stopp-sõnade loendit, millega eemaldatakse sagedasemad mitteolulised sõnad, nagu enamik asesõnu ja sidesõnu ning *olema* vormid. Seejärel saab kasutada mõnda statistilist meetodit, et sõnadevahelise seose tugevust mõõta. (Muischnek & Kaalep 2009: 163–164) Samuti saab kasutada näiteks (tähenduslikku)

kompositsioonilisust mõõtvaid mudeleid ning asendatavuse katseid, kus vaadatakse, kas ühendi sellist varianti, kus mõni sõna on asendatud, mõni sõna on lisatud või see on veidi teisiti sõnastatud, esineb ka korpuses, mispuhul on selle püsiühendiks olemise tõenäosus väiksem (Ramisch 2023: 39–41).

Märgendamise etapi jaoks on oluline määratleda, kuidas püsiühendid tekstis käituvad ja kuidas see eri liiki püsiühendite vahel varieeruda võib. Esiteks näiteks see, kui palju ja kuidas see võib erineda leksikonis esinevast vormist, võrreldes näiteks *võta näpust* ja *joonde ajama* varieerumisvõimalusi, ei varieeru esimene neist üldse, teises võib tegusõna vormiliselt varieeruda, aga käändsõna on muutumatu. Teiseks sõnade järjestus ja kui kaugel need üksteisest olla võivad, eriti kuna need võivad vahel ka eri osalausetes olla, nagu ühend *abi vajama* lauses „Naine nendib, et sai haiglast abi, mida vajas.“ Märgendamisprogrammi lihtsustamiseks võib tuvastamisetapi põhjal koostatavas leksikonis märkida iga püsiühendi juurde, kui palju ja kuidas see muutuda võib, mille lihtsustamiseks võib omakorda kasutada samamoodi käituvatest ühenditest koosnevaid klasse, näiteks ühendverbid, mille mitte-verbiline komponent ei saa muutuda, ja tugiverbiühendid, mille mitte-verbiline komponent võib muutuda arvus ja objektikäänetes. (Muischnek & Kaalep 2009: 164–168)

Järgmisena tuleb märgendamisel tuvastada, kas leitud sõnad, mis võivad kuuluda püsiühendisse, moodustavad ka selles kontekstis ühendi või on juhuslikult kokku sattunud. Selleks saab kasutada märgendamise alguses tehtud morfoloogilist ühestamist, et kontrollida leitud sõnade sõnaliike, ja süntaktilist analüüsi, et kontrollida nende lauseliikmeid. Samuti tuleb vahet teha idiomaatilisel ja otsesel kasutusel, näiteks ühend *solvangut alla neelama* kasutab *alla neelama* idiomaatilise tähendusega, *tabletti alla neelama* kasutab seda aga otsese tähendusega. Selleks on pakutud eri meetodeid, näiteks süntaktilise jäikusega arvestamine, mille puhul võib kontrollida nimisõna laiendiga esinemise võimalikkust, või kontekstiga arvestamine, mille puhul kontrollitakse, kas samas (osa)lauses esineb otsese tähendusega tihti koos esinevaid sõnu või süntaktilisi seoseid. (Muischnek & Kaalep 2009: 168–169) Lisaks on kasutatud sõnavektoritel põhinevaid lähenemisi, et mõõta püsiühendite kompositsionaalsust nende ja nende komponentide esinemiskontekstide sarnasuste põhjal (Ramisch 2023: 72–75).

Püsiühendite automaatse töötlemise arendamisel on olulised ressursid peale leksikonide ka korpused, mida kasutatakse püsiühendite tuvastamise tööriistade treenimiseks ja testimiseks (Ramisch 2023: 24–26). Korpuste märgendamisel on mainitud analüüsimismeetodeid oluline teada, sest need annavad aimu, mida püsiühendite (käsitsi) märgendamisel silmas pidada. Kui märgendamisel märkida üles, mis teeb sellest ühendist püsiühendi (nt UniDive'i puhul püsiühendi tuvastanud testi abil), siis saab ka teada, millistele meetoditele rohkem tähelepanu

pöörata. Lisaks aitab märgendamine näha, kuidas eri liiki püsiühendid varieeruda võivad. See kõik näitab korpuste märgendamise olulisust.

3. Püsiühendite märgendamine korpustes

3.1. UniDive'i eellased

UniDive'ile lähedane initsiatiiv PARSEME[2] on rahvusvaheline kogukond, mis arendab mitmekeelseid (tegusõnaliste) püsiühenditega märgendatud korpuseid ja mis sai alguse 2013–2017 toimunud homonüümsest COST (*European Cooperation in Science and Technology*) *action*'ist (PARSEME 2024), mis keskendus püsiühendite rollile süntaksianalüüsis. PARSEME nimi tähendab *Parsing* (süntaksianalüüs) *and Multi-word Expressions* (Savary jt 2015: 1).

Kuni 2020. aastani oli PARSEME kogukonna üks peamisi tegevusi organiseerida ja pakkuda materjale PARSEME *shared tasks* (jagatud ülesannete) jaoks, mille eesmärk on omakorda pakkuda raamistikku automaatselt püsiühendite tuvastamist arendavatele gruppidele (PARSEME 2024). Alates 2021. aastast on initsiatiivis tegeletud korpuste arendamisega ja uute keelte korpuste märgendamisega, 2023. aastal välja antud 1.3 versioon sisaldas näiteks 26 keele korpuseid (Savary, Ben Khelil, jt 2023: 24). PARSEME on keskendunud peamiselt tegusõnalistele püsiühenditele, aga alates 2021. aastast on tegelenud ka teist liiki püsiühendite käsitluste arendamisega (PARSEME 2024).

Teine sarnase üldeesmärgiga, aga teise keskendumiskohaga rahvusvaheline initsiatiiv on *Universal Dependencies*[3] (UD), millega seoses tehakse keelteülest morfoloogilist märgendust ja sõltuvussüntaktilist märgendust, sh märgendatakse sõnaliike, morfoloogilisi kategooriaid, süntaktilisi funktsioone ja lauseliikmete vahelisi sõltuvussuhteid (De Marneffe jt 2021: 1). Sellega seoses tehakse puudepanki ehk keelekorpuseid, mis koosnevad süntaktiliselt märgendatud lausetest ja mida on UD-ga seoses valminud üle 150 keeles kokku üle 200, eesti keeles näiteks kaks (UD *s. a.*).

3.2. UniDive'i initsiatiiv ja motivatsioon

Nii UD kui ka PARSEME initsiatiivi eesmärk on keelteülese märgendusskeemi väljatöötamine, selle järgi korpuste märgendamine ja nende korpuste kasutamine loomuliku keele töötluse (ingl *Natural Language Processing* ehk NLP) tehnoloogia arendamiseks, et neid automaatsemalt märgendada. Need initsiatiivid on peamiselt tegutsenud eraldi ning nad kasutavad ka veidi erinevaid termineid ja meetodeid, aga 2022. aastal alustati uut COST *action*

[2] https://gitlab.com/parseme/corpora/-/wikis/home.
[3] https://universaldependencies.org/.

projekti nimega UniDive, mille üks eesmärk on nende kahe initsiatiivi ühendamine. „UniDive“ on lühend nimest *Universality, Diversity and Idiosyncrasy* (mitte-reeglipärasus) *in Language Technology*. (Savary, Stymne, jt 2023: 1–2)

UniDive'i initsiatiiv on seega samuti COST *action* projekt, kestusega 2022—2026, ja selle üldine eesmärk on kooskõlastada keeleline mitmekesisus keeletehnoloogia kiire arenguga. UniDive'i jaoks on oluline keeltevaheline ja -sisene mitmekesisus. Sellel on neli töörühma: esimene tegeleb korpuste märgendamisega, teine leksikoni-korpuse liidesega, kolmas mitmekeelse ja keelteülese keeletehnoloogiaga ning neljas mitmekesisuse mõõtmise ja edendamisega. Igal töörühmal on omad alamülesanded ja esimese töörühma ülesanne 1.2 ongi märgendada püsiühendeid eri keelte üleselt ühtlasena ja arendada rahvusvahelise koostöö jaoks ühtne raamistik, ühendades selles osas UD ja PARSEME initsiatiive. (UniDive *s. a.*)

UniDive'iga on seotud erineval määral palju riike, näiteks korralduskommitees on 37 riiki, vähemalt ühe töörühmaga on seotud 48 riiki (arvestades ka Kosovot ja Hong Kongi) ja esimese töörühmaga ehk korpuste märgendamisega on seotud 43 riiki (arvestades ka Kosovot). Kuigi suures osas moodustavad töörühmade koosseisu Euroopa riigid, on huvlisi ka kaugemalt, näiteks India on seotud kolme töörühmaga ja Iraan ainult esimesega. (COST *s. a.*)

3.3. Püsiühendid UniDive'i initsiatiivi raamistikus

UniDive'i initsiatiivis on sõnastatud mõiste *multiword expression*, mida on eesti keelde tõlgitud püsiühendina (Muischnek & Kaalep 2009: 157) ja millest selles töös lähtun – see on (järjestikune või mitte-järjestikune) sõnade ühend, mis:

1) kaldub millegi, näiteks süntaksi või semantika, poolest keele üldistest reeglitest kõrvale,
2) koosneb vähemalt kahest sõnast, millest üks on süntaktilises analüüsis peasõna ja teine/teised on laiendid, ning
3) sisaldab vähemalt kahte komponenti, mis esinevad alati samade lemmadena (Savary jt *s. a.* (a): 2).

Viimast punkti on vaja välja tuua seepärast, et osades püsiühendites on ka mõni nn lünk (ingl *open slot*), mis ei ole kindel sõna, aga peab püsiühendi tekstis kasutamisel kindlate grammatiliste tunnustega sõnavormiga täidetud olema (Lexicalized ... *s. a.*), näiteks *[kellegi] käe all*. Kõiki kolme punkti illustreerib hästi püsiühend *südamest tänulik*. Esiteks tähendab see 'väga tänulik', aga seestütlevas käändes nimisõna ei kasutata regulaarselt omadussõna intensiivistajana ega üldse omadussõna laiendina. Seega on antud sõnapaar nii semantiline kui ka süntaktiline anomaalia (loetelu esimene punkt). Teiseks on sellel peasõna *tänulik* ja laiend

südamest ning kolmandaks sisaldab see kahte samade lemmadena esinevat osa, mis jäävad samaks ka ühendi varieerudes, näiteks vormides *südamest tänulikule* ja *südamest tänulikke*.

Lisaks on UniDive'i püsiühendite määratluse aluseks hüpotees, et püsiühenditele on omane teatav semantiline mittekompositsioonilisus, mis toob kaasa piiratud paindlikkuse. Seega ei tuvastata neid püsiühendite testides mitte selle järgi, kas püsiühendi tähendus on kompositsiooniline, sest see moodus ei pruugi olla selge ja objektiivne, vaid kasutatakse paindumatuse printsiipi. Niisiis tuleb märgendamisel järjest teste läbi proovida, kuni mõni määrab ühendi püsiühendiks, ja kui ükski seda ei tee, siis pole selle käsitluse järgi tegu püsiühendiga. Testid on näiteks järgmised: „Kas reeglipärane morfoloogiline muutus toob kaasa ootamatu tähendusnihke?" (näiteks ei saa öelda *kunstielud*) ja „Kas kandidaadi koordinatsioon teise sama põhisõnaga kandidaadiga toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?" (näiteks on kummaline *vaate- ja ekraanipilt*). (Savary jt *s. a.* (a): 2, 8)

UniDive'i raamistikus loetakse püsiühenditeks ka liitsõnu, sest nende kokku või lahku kirjutamise reeglid varieeruvad eri keeltes, näiteks inglise keele *cold weapon* ja eesti keele *külmrelv* on sama tähendusega, aga erinevalt kirjutatud. Kohati on need reeglid kokkuleppelised ka ühe keele sees, näiteks *mustkunst* ja *must turg*. (Kull 1967: 47) Kui tegu on rohkem kui kahest sõnast koosneva liitsõnaga, millest ainult üks osa on püsiühend, siis on UniDive'is pakutud, et võiks märgendada ainult vastavat osa liitsõnast (Savary jt *s. a.* (a): 4). Näiteks *muusikamaailma* puhul on *maailm* püsiühend, aga *muusika* ja *maailm* ei moodusta koos püsiühendit. Samuti võib liitsõna koosneda kahest püsiühendist, näiteks liitsõnas *ajaloomaailm* on püsiühendid *ajalugu* ja *maailm*. Siiani pole seda pakkumist ametlikult rakendatud, mistõttu soovitatakse märgendada ikkagi tervet liitsõna püsiühendina, isegi kui püsiühend on ainult osa liitsõnast (Savary jt *s. a.* (a): 4), näiteks *muusikamaailma* puhul märgendada terve sõna, mitte ainult *maailma* osa. See soovitus võib tuleneda sellest, et osades märgendamisprogrammides pole selliselt märgendamine võimalik või see võib ehk analüüsi raskendada. Antud töös olen märgendanud selle mitteametliku pakkumise järgi ehk ainult püsiühendi osa liitsõnast, kuna kasutatud märgenduskeskkonnad on seda võimaldanud ja see annab lemmadest parema ülevaate.

Lisaks on UniDive'is käsitletud püsiühendite kategooria äärealasid. Projektis kaasatakse püsiühendilisi terminoloogilisi üksuseid, näiteks *mõõtkava* ja *läbimõõt.* Nende puhul ei pruugi tähendused olla valdkonnavälisele inimesele üldtuntud ning tähenduste leidmine on eriti keeruline siis, kui neid pole ka sõnaraamatutes, näiteks *roheline bioloogia* ja *valge bioloogia*, millest esimene uurib pigem suuremaid elusorganisme, nagu loomi ja taimi, ja teine

väiksemaid, nagu rakud ja molekulid. UniDive'is ei loeta püsiühenditeks pärisnimesid, näiteks *Valge Maja*, kusjuures pärisnimede tuvastamiseks (ja kõrvale jätmiseks) on ka omaette testid. (Savary jt *s. a.* (a): 4) Arutletud on ka selle üle, kas UniDive'i projekti raames lugeda püsiühendite alla võrdlusi, näiteks *nagu kurjast vaimust vaevatud*, aga kuna selle osas pole veel otsusele jõutud, (Savary jt *s. a.* (a): 12) jätan need ka antud tööst välja.

Kui võrrelda UniDive'i initsiatiivi definitsiooni esimeses peatükis välja toodud püsiühendite tunnustega, siis on selles täpsustatud püsiühendi elementide olemust, arvu ja et need ei pea olema järjest. Samuti on arvestatud teatud varieeruvusega, näiteks lünkade ja eri testide näol, ning oluline rõhk on mitte-kompositsioonilisusest tuleneval piiratud paindlikkusel ja idiosünkraatilisusel. Lisaks on täpsustatud, et ühendid, mis on idiosünkraatilised ainult sageduse poolest, ei kuulu püsiühendi definitsiooni sisse, mis ehk katab teatul määral sageduse tunnuse. (Savary jt *s. a.* (a): 1–2, 4)

4. Püsiühendite liigituse ja testimise katsetamine eesti keele materjalil

UniDive'is liigitatakse püsiühendeid nende lausedistributsiooni järgi neljaks: tegusõna-, nimisõna-, modifitseerijafraasina (omadussõna- või määrsõnafraasina) ja funktsiooni fraasina, nagu kategoriseeritakse selles mittelauseliikmelisi fraase (näiteks sidesõna- või hüüdsõnafraasid), kasutatavad püsiühendid (Savary jt *s. a.* (a): 7). Selle töö raames märgendan nimisõna- või modifitseerijafraasina kasutatavaid püsiühendeid. Nimisõna distributsiooniga püsiühendid jagunevad asesõnalisteks, nimisõnalisteks ja deverbaalseteks nimisõnalisteks püsiühenditeks (Savary jt *s. a.* (a): 8–9). Näiteks *igaüks* on asesõnaline, *vikerkaar* on nimisõnaline ja *ülevaade* on nominalisatsioon verbilisest püsiühendist *üle vaatama*. Modifitseerija distributsiooniga püsiühendid jagunevad omadussõnalisteks, määrsõnalisteks ja deverbaalseteks modifitseerijana kasutatavateks püsiühenditeks (Savary jt *s. a.* (a): 7). Näiteks *surmani väsinud* on omadussõnaline, *käe-jala juures* on määrsõnaline ja *läbilõikav* on deverbaalne.

UniDive'i raamistikus on nominaalsete ja modifitseerija funktsioonis püsiühendite märgendamisel kolm etappi:

1) tuvastada püsiühendikandidaat (tugineb suuresti märgendaja lingvistilisele teadmisele ja keelevaistule),
2) teha kindlaks kandidaadi leksikaliseerunud osad ehk kohustuslikud komponendid, kusjuures neid peab olema vähemalt kaks, ning
3) rakendada kandidaadile järjest vajalikud testid ja otsustada nende põhjal, kas tegu on püsiühendiga.

Nominaalsete püsiühendite jaoks on 15 testi, millest esimesed kaks tuvastavad, kas tegu on asesõnalise püsiühendiga, järgmised kolm kontrollivad, et tegu poleks pärisnimega, ja viimane ehk viieteistkümnes test tuvastab, kas tegu on nominaliseeritud verbilise püsiühendiga. Seega üheksa testi ehk testid 6–14 saavad määrata püsiühendi nimisõnaliseks. Modifitseerija funktsioonis püsiühendite jaoks on 8 testi, millest üks tuvastab, kas tegu on deverbaalse modifitseerija kujul püsiühendiga, ja ülejäänud saavad määrata püsiühendi kas omadussõnaliseks või määrsõnaliseks.

4.1. Nimisõnafraasiliste püsiühendite testid

Testid on mõeldud rakendamiseks kindlas järjekorras, kuni mõni määrab püsiühendikandidaadi püsiühendiks. Testide esialgne järjekord ja nimed on võetud UniDive'i

juhendist (Savary jt *s. a.* (a): 8–10). Testide rakendamine käib nii, et kui vastus küsimusele on jah, siis tuleb ühend püsiühendiks märgendada, ja kui vastus on ei, siis tuleb liikuda järgmise testi juurde. Sellele on eranditeks testid 3, 4 ja 5. Vastavalt UniDive'i püsiühendi definitsioonile kontrollivad testid eri tüüpi idiosünkraatilisust, näiteks morfoloogilist, süntaktilist või semantilist, ja sellest tulenevat piiratud paindlikkust.

Käesolev testide järjekord oli UniDive'i juhendis 2024. aasta sügisel, mil märgendama asusin. Vahepeal on veidi muudetud mõne testi järjekorda ja sõnastust. Täpsemalt on vahetatud 11. ja 12. testi ehk süntaktilise muutuse ja koordinatsiooni testi järjekorda. (Savary jt *s. a.* (b): 12) Kuna olin muutuse ajaks juba palju märgendanud, siis kasutasin sügist versiooni edasi, et ei peaks nii palju ümber märgendama.

1) [IS_PRON] – Kas püsiühendikandidaat on lauses asesõna rollis?
2) [ON_PRON_LIST] – Kas kandidaat on asesõnaliste püsiühendite suletud nimekirjas või kas selle peaks sinna lisama?

   Asesõnade listi test pole formaliseeritud ja eesti keelele pole asesõnaliste püsiühendite nimekirja loodud, nii et olen asesõnalistele püsiühendikandidaatidele rakendanud ka ülejäänud testid ja positiivse vastuse puhul testiks märkinud esimese testi. Asesõnalised püsiühendid on näiteks *igaüks* ja *teineteise*.
3) [SPECIFIC_REF] – Kas kandidaati kasutatakse selles kontekstis diskursuse ühele kindlale entiteedile viitamiseks?

   Kui jah, näiteks *tädi Maali* lauses *Isa läks Elvasse tädi Maalile külla*, kus juttu on konkreetselt kellegi tädist, siis tuleb liikuda 4. testi juurde. Kui ei, näiteks *tädi Maali* lauses *Sellist juttu räägivad kaks tädi Maalit kuskil kohvikus*, kus juttu on üldiselt teatud tüüpi inimestest, siis tuleb liikuda 6. testi juurde.
4) [CONCEPT_NAMING_CONV] – Kas kandidaat viitab kõigile entiteedi esinemisjuhtumitele ehk kas kandidaat saab viidata ka mõnele muule entiteedile, millel on samad omadused kui selles kontekstis viidatud entiteedil, ilma et kandidaati peaks kuidagi muutma?

   Kui jah, siis tuleb liikuda 5. testi juurde. Kui ei, aga selle tõttu, et ei saa olla teist sarnast entiteeti, näiteks *Suur Pauk* või *Kadaka Selver*, sest on ainult üks sellise nimega pood, siis tuleb liikuda 5. testi juurde. Kui lihtsalt ei, siis on ilmselt tegu pärisnimega, mitte püsiühendiga, ja testimine tuleb lõpetada.
5) [SEM_TYPE] – Kas kandidaadi viidatud entiteet on isik, organisatsioon, asukoht, inimtoode või sündmus?

Kui jah, näiteks *lihavõtted* või *Kadaka Selver*, siis on ilmselt tegu pärisnimega ja testimine tuleb lõpetada. Kui ei, siis tuleb liikuda 6. testi juurde.

6) [CRAN] – Kas kandidaat sisaldab jäänukmorfi (ingl cranberry word)?

   Jäänukmorf, näiteks *viker* sõnas *vikerkaar* ja *maan* sõnas *maantee*, või -sõna, näiteks *pilkane* ühendites *pilkane pimedus/öö*, tähendab sõna või morfi, mis esineb ainult ühes või paaris püsiväljendis. See test testib seega leksikaalset idiosünkraatilisust ning selle testimiseks saab otsida näiteks Sõnaveebist ühendi komponente eraldi ja uurida ühendi etümoloogiat.

7) [MORPH] – Kas korrapärane morfoloogiline muutus toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

   Vormimuutmise test testib morfoloogilist idiosünkraatilisust ja selle testimiseks võib proovida ühendit näiteks mitmusesse panna. Näiteks ei saa liitsõna *kunstielu* panna mitmusesse kui *kunstielud* ja ühendit *kimpsud-kompsud* ainsusesse kui *kimps-komps*, mida võib võrrelda näiteks liitsõnaga *unenägu* : *unenäod*, mille puhul saab mõlemat vormi kasutada. Osade püsiühendite puhul ei saa seda testi rakendada semantiliste piirangute tõttu, näiteks ühendit *hambad ristis* ei saa muuta kujule *hammas ristis* juba sellepärast, et üksik asi ei saa risti olla. Samuti ei saa seda rakendada ühenditele, mille põhisõna ei saa ka üksinda olla mitmuses või ainsuses, näiteks *kirjandus*.

8) [IRREG_STRUCT] – Kas kandidaadil on ebatavaline sisemine süntaktiline struktuur?

   Ebatavalise struktuuri test testib süntaktilist idiosünkraatilisust ja testimiseks võib proovida leida regulaarse sisemise struktuuriga varianti, näiteks *eite-taati* puhul *eit ja taat*. Kuigi sarnaselt ühendile *eite-taati* saab öelda ka näiteks *tüdrukut-poissi* või *kassi-koera*, siis saab *eite-taati* kasutada ka nominatiivsena, teisi aga ainult partitiivi vormina.

9) [IRREG_STRUCT_DISTRIB] – Kas kandidaadil on selle lausedistributsiooni kohta ebatavaline sisemine struktuur ehk kas sellel pole nimisõnafraasi struktuuri?

   Ebatavaliste sisemiste lauseliikmete test kontrollib samuti idiosünkraatilisust. Näiteks võib tuua *meelespea* ja *ära-mind-unusta*, millel on sisemine verbifraasi struktuur, aga mis käituvad nimisõnana.

10) [INSERT] – Kas tavaline täiendite (nt omadussõnade, määrsõnade, asesõnade, arvsõnade, kaassõnafraaside või relatiivlausete) lisamine mõnele kandidaadi komponendile toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

    Ka sõna lisamise test testib idiosünkraatilisust ja selle testimisviis on juba küsimuses endas olemas. Selle iseloomustamiseks võib näiteks tuua võrdluse liitsõnade *lasteaed* ja *lasteraamat* vahel, sest fraas *väikeste laste raamat* säilitab liitsõna algse tähenduse,

aga *väikeste laste aed* mitte. Samas on osade ühendite puhul, mille esiosa on nominatiivis, raske leida täiendit, mis ei muudaks täiendatavat sõna. Näiteks et lisada täiend sõnale *okas* liitsõnas *okastraat*, peab selle panema osastavasse – *terava okka traat*.

11) [SYNT] – Kas korrapärane süntaktiline muutus toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

    Süntaktilise muutmise test kontrollib samuti süntaktilist idiosünkraatilisust, aga seda saab katsetada lahku kirjutatavate ühendite peal. Näiteks ei saa ühendi *tuule peal* asemel öelda *tuulel* ning ühendi *maa ja ilm* asemel *ilm ja maa*.

12) [COORD] – Kas kandidaadi koordinatsioon teise sama põhisõnaga kandidaadiga toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

    Koordinatsiooni test testib semantilist idiosünkraatilisust ja testimisviis on küsimuses välja toodud. Positiivseks näiteks võib tuua ühendi *vaatepildi* puhul tähendusliku nihkega *vaate-* ja *ekraanipilt*, negatiivseks näiteks võib tuua *kirsi-* ja *õunapuud*.

13) [ID] kas kandidaadi põhisõna on kandidaadi hüperonüüm, mida saab ümber sõnastada kui „[kandidaat] on teatud tüüpi [põhisõna]“?

    Hüperonüümsuse test kontrollib semantilist idiosünkraatilisust ja testimisviis on küsimuses sõnastatud. Näiteks *unenägu* ei ole teatud tüüpi nägu, aga *elurõõm* on teatud tüüpi rõõm.

14) [LEX] – Kas kandidaadi ühe leksikaalse komponendi korrapärane asendamine seotud sõnadega, mis on suhteliselt suurest semantilisest klassist, toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

    Sõna asendamise test testib leksikaalset idiosünkraatilisust ja testimisviis on küsimuses olemas. Näiteks kaubamaja ei saa ümber sõnastada kui *tootemaja* või *kaubahoone*. Mõnede püsiühendite puhul on seotud sõnaga variant olemas, aga hoopis teise tähendusega, näiteks *välismaa* ’mingist riigist väljaspool olev maa või riik’ ja *sisemaa* ’merest või rannikust eemal olev maismaaosa'. Selle testi juures võib probleemiks tulla sobiva seotud sõna leidmine, näiteks kas *süvamuusika* tähendusliku nihkega ümbersõnastuseks sobib *pinnamuusika*. Nagu näha, on osad testid siiski veidi subjektiivsed, kuigi püüeldakse objektiivsuse poole.

15) [DEVERBAL] – Kas kandidaat sisaldab deverbaalset nimisõna ja kas seda saab ümber sõnastada (samas kontekstis) kasutades tegusõnafraasi, mis on läbinud mõne verbilise püsiühendi testi?

Verbiliste püsiühendite testid on eelnevalt PARSEME initsiatiivi raames välja töötatud (PARSEME s.a.), aga eesti keelele kohandamata. Deverbaalsuse test pole seotud niivõrd püsiühendilisuse testimisega, kuivõrd leitud püsiühendi liigi määramisega. Seega, kui vastus küsimusele on positiivne, siis on tegu nominaliseeritud verbilise püsiühendiga, mitte nimisõnafraasilise püsiühendiga, ja testiks läheb kirja see 15. test, isegi kui ka mõne eelmise testi tulemus oli positiivnes. Näiteks *ettepanek* on nominalisatsioon verbiühendist *ette panema*.

Asesõnaliste püsiühendite puhul peaks tulevikus kas tegema eestikeelsete asesõnaliste püsiühendite nimekirja, mida saaks teises testis kasutada, või käsitlema nimi- ja asesõnalisi püsiühendeid nagu modifitseerijafraasilisi, kus püsiühendi täpsem liik määratakse peale mõnelt testilt positiivse tulemuse saamist. Viimasel juhul saaks ka teada, milline test asesõnalise püsiühendi tuvastas, aga kuna asesõnu on piiratud arv, siis on nimekirja tegemine ilmselt mõistlikum.

4.2. Modifitseerijafraasiliste püsiühendite testid

Modifitseerijafraasiliste püsiühendite juures on omadus- ja määrsõnafraasilistel püsiühenditel samad testid ja alles peale testi positiivset tulemust määratakse need erinevatesse liikidesse. UniDive'i eeskujul jätsin välja mõned nimisõnafraasiliste püsiühendite testid, nagu COORD ja ID test, sest neid ei saa omadus- ja määrsõnafraasiliste püsiühendite puhul eriti testida.

Märgendamise jooksul lisasin SYNT testi, mis oli algselt koos COORD ja ID testidega välja jäetud, sest leidsin sellele kaks näidet, nimelt *risti ja põiki* ning *ikka ja jälle.* SYNT testi lisamisega asendasin algselt neljanda testi IRREG_STRUCT_DISTRIB, sest ei arvanud, et leian ühtegi tollelt testilt positiivset tulemust saavat püsiühendit. Siiski leidsin ebatavaliste sisesmiste lauseliikmete testile hiljem mõned head näited, nagu *nii-öelda* ja *võib-olla*. Seega otsustasin selle kaheksanda ehk viimase testina tagasi lisada ja märgendasin leitud püsiühendid selle testiga järjekorrast hoolimata, nagu tegin ka deverbaalsuse testiga. Sellise märgendamise järjekorra kasuks räägib ka see, et kõik sellelt testilt positiivse tulemuse saanud püsiühendid saaksid juba vormimuutmise testi juures positiivse tulemuse, sest ei saa öelda näiteks *nii-öeldud* ega *võis-olla*.

Ka DEVERBAL-MOD test on üks neist, mida sügisel juhises ei olnud ja mis hiljem juurde lisati. Märgendamise jooksul oli mul korduvalt ette sattunud partitsiipe, mis süntaktiliselt käitusid omadussõnadena ja esinesid vahel isegi käänatult, aga olid sisuliselt ikkagi tegusõnad,

mistõttu tekkis nende märgendamisel dilemma, kas neid märgendada või mitte. Selle küsimuse lahendas mugavalt deverbaalsuse testi lisamine analüüsi, eriti kuna selliste püsiühendite mõõduka sageduse tõttu ei nõudnud üle märgendamine nii palju lisatööd. Siiski on olemas mõned püsiühendilised partitsiibid, millel pole vastavat verbivormi olemas, näiteks *kõrvulukustav* ja *kõrvulukustavalt*, sest ühend *kõrvu lukustama* ei tähenda midagi. Neid märkisin seega lihtsalt modifitseerijafraasilisteks püsiühenditeks, sest need pole verbiühenditest tuletatud.

1) [CRAN] - Kas kandidaat sisaldab jäänukmorfi või -sõna?

   Näiteks *algu* esineb ainult püsiühendis *esialgu*.

2) [MORPH] – Kas korrapärane morfoloogiline muutus toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

   Näiteks püsiühend *kontideni lõikav* tähendab midagi muud kui *kondini lõikav*. Seda testi saab teha ainult selliste omadus- ja määrsõnade puhul, mis sisaldavad käändsõnu, näiteks *ümberringi* puhul on mõlemad komponendid muutumatud sõnad ja seda testi ei saa sellele sõnale rakendada.

3) [IRREG_STRUCT] – Kas kandidaadil on ebatavaline sisemine süntaktiline struktuur?

   Näiteks *käe-jala juures* ja *eelkõige*, regulaarne oleks ehk *käe ja jala juures* ning *kõige eel*.

4) [SYNT] – Kas korrapärane süntaktiline muutus toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

   Näiteks *risti ja põiki* puhul ei saa öelda *põiki ja risti*.

5) [INSERT] – Kas tavaline laiendite (nt omadussõnade, määrsõnade, asesõnade, arvsõnade, kaassõnafraaside või relatiivlausete) lisamine mõnele kandidaadi komponendile toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

   Näiteks *ükskord* ei ole sama kui *ainult üks kord* ning *hambad ristis* ei ole sama kui *valged hambad ristis*.

6) [LEX] – Kas kandidaadi ühe leksikaalse komponendi korrapärane asendamine tähenduslikult sarnaste sõnadega, mis on suhteliselt suurest semantilisest klassist, toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?

   Näiteks püsiühendit *ükskõikne* ei saa ümber sõnastada kui *kakskõikne* ja püsiühendit *peaaegu* kui *ligiaegu*.

7) [DEVERBAL-MOD] Kas kandidaat sisaldab deverbaalset modifitseerijat ja kas seda saab ümber sõnastada (samas kontekstis) kasutades tegusõnafraasi, mis on läbinud mõne verbilise püsiühendi testi.

Kui jah, siis tuleks see määrata modifitseerijafraasiliste püsiühendite verbiliste püsiühendite allrühma ModMWE.VID (ingl *a ModMWE with the VMWE sub-class*). Näiteks *ettevaatlik* ja *arupidavalt*.

8) [IRREG_STRUCT_DISTRIB] – Kas kandidaadil on selle lausedistributsiooni kohta ebatavaline sisemine struktuur ehk kas sellel pole modifitseerijafraasi struktuuri?

   Näiteks *nii-öelda*, mis käitub määrsõnana, aga mille sisemine struktuur on adverb-verb, ja *võib-olla*, mis käitub määrsõnana, aga mis on verbiühendi struktuuriga.

Tulevikus oleks mõistlik nii nimisõna- kui ka modifitseerijafraasiliste püsiühendite testide järjekorda loogilisemaks muuta, eriti võiks nihutada nimekirja algusesse deverbaalsuse testid, sest need määravad püsiühendi liiki nagu ka asesõnade testid nimisõnafraasiliste püsiühendite testide alguses. Samas oleks selleks vaja kohandada verbifraasiliste püsiühendite testid eesti keelele. Samuti võiks varasemaks tõsta ebatavaliste sisemiste lauseliikmete testi, mis võiks eelneda vormimuutuse testile varem mainitud põhjuste tõttu.

5. Materjal ja meetod

Uurimismaterjalina kasutan esindusvalimiga valitud tekste eesti keele Universal Dependencies (UD) 2.14 puudepangast (Zeman jt 2024) – keelekorpusest, mis koosneb süntaktiliselt märgendatud lausetest. Eesti keele UD puudepangas on umbes 400 000 sõna ja tekstid on jaotatud ilukirjanduslikeks, ajakirjanduslikeks ja (populaar)teaduslikeks tekstideks (Muischnek, Müürisepp & Puolakainen 2014: 111). Valisin selle korpuse, sest see oli juhendaja kaudu mulle kättesaadav, lisaks on selles juba märgendatud lemmad, mille järgi saab eristada liitsõnu, ning failid on jaotatud tekstiliigi järgi. Samuti sain selles märgendada sidusaid tekste üksikute lausete asemel, mida oleks ehk mõne muu korpusega tegema pidanud. See aitas mul püsiühendite kontekstist paremini aru saada, mis oli eriti kasulik mulle võõraste sõnaühendite puhul.

Tekste valisin esindusvalimi põhjal, et igast tekstiliigist oleks sama palju lauseid. Märgendasin 500 lauset iga tekstiliigi kohta, kusjuures sõnesid ehk tekstisõnasid oli kõige rohkem (üle 6300) populaarteaduslikes tekstides ja kõige vähem (siiski üle 5600) ilukirjanduslikes tekstides, ilmselt olid esimeses pikemad laused. Kokku on minu märgendatud korpuses 1500 lauset ja 18 164 sõnet, mille hulgas pole arvestatud kirjavahemärke, küll aga on arvestatud arve.

Ilukirjanduslikest tekstidest märgendasin kolme autori tekste, ajakirjanduslikest tekstidest nelja ajalehe tekste ja populaarteaduslikest tekstidest kaks juppi ajakirjast Horisont. Valisin märgendamiseks populaarteaduslikud tekstid (ajakirjast Horisont), mitte teaduslikud tekstid, sest teaduslikes tekstides on rohkem terminoloogiat, millega enda kurssi viimine oleks liiga ajamahukas. Failinimedest on näha, et tekstid on aastastest 1999–2007.

Alguses kasutasin märgendamiseks Microsoft Wordi dokumente UniDive'i initsiatiivi näitemärgenduse eeskujul. Selle järgi peaks tekstis igale tuvastatud püsiühendile lisama kommentaarina selle tuvastanud testi ja värvima seda vastavalt selle liigile, näiteks asesõnalised püsiühendid on punased ja nimisõnalised helesinised. Kopeerisin teksti faili nii, et iga lause oli järjepidamise lihtsustamiseks loetelus eraldi real, ja lisasin iga tuvastatud püsiühendi juurde kommentaarina lisaks selle tuvastanud testi ka näite põhjusest, miks see testist läbi ei saanud. Näiteks testi NMWE.7 MORPH „Kas korrapärane morfoloogiline muutus toob kaasa ootamatu tähendusnihke?" puhul oleks näiteks *kunstielud* ja testi NMWE.12 COORD „Kas kandidaadi koordinatsioon teise sama põhisõnaga kandidaadiga toob kaasa ebagrammatilisuse või ootamatu tähendusnihke?" puhul oleks näiteks *vaate- ja ekraanipilt*.

Samuti märkisin üles kahtlema panevad kohad, et neid hiljem põhjalikumalt analüüsida. Märgendamisprotsessile lähenesin nii, et igas kümnest lausest koosnevas jupis tõmbasin alguses liitsõnadele jooned alla, sest enamik püsiühendeid on liitsõnad, ja panin püsiühendikandidaadid paksu kirja. Seejärel tegin kandidaatidega testid läbi ja peale kommentaari lisamist tegin leitud püsiühendid värviliseks. Lisaks arutasin juhendajaga kahtlema panevate kohtade üle, näiteks *Mõeldud-tehtud*, mille määrasime eraldi lauseks ja seega fraseologismiks, mitte püsiühendiks.

Hiljem hakkasin märgendama INCEpTIONi tekstimärgendamise programmis (Klie jt 2018), et lihtsustada edasist andmeanalüüsi, kuna seal on edasiseks töötlemiseks mugavamad failivormid. Sinna saab teha projektidena korpused, kus saab üles laetud tekstifaile mugavalt märgendada eri kihtide peal. INCEpTIONisse tõstsin algsed UD puudepanga failid, mis olid CONLLU-formaadis. Süntaksianalüüsi jaoks on puudepanga tekstidesse taastatud elliptilised ehk väljajättelised sõnad, näiteks „23. veebruaril sõidetakse 20 km, 26. veebruaril [sõidetakse] 15 km“, mille pidin enne failide üleslaadimist eemaldama.

Märgendamiseks tegin INCEpTIONis uue kihi, kuhu lisasin väljad „tüüp“, „test“, „näide“ ja „kommentaar“. Ekraanipilti INCEpTIONist vt Lisast 1. Esimesel väljal saab valida ühendi tüübi: AdjID, AdvID, NID, PronID, VMWENom, ModMWE.VID, Muu ja Pole[4], viimased kaks kahtlema panevate kohtade märkimiseks. Teises väljas saab valida testi, mis püsiühendi tuvastas, ja kolmandas lisada näite sellest, miks see testist läbi ei saanud. Kuna algsed UD puudepanga failid on CONLLU-formaadis, mis võimaldab märkida sõltuvussuhteid, siis sain lisada lemmat näitavale kihile värvifiltri, mis kuvaks teise värviga liitsõnad ehk lemmad, mis sisaldavad „_“, näiteks „maa_ilm“. Märgendatud failid tõmbasin alla formaadis „WebAnno TSV v3.3 (WebAnno v3.x)“. Ekraanipilti allalaetud tsv-failist vt Lisast 2. INCEpTIONi projekti kokkupakitud kaust on saadaval OneDrive’is (https://tartuulikool-my.sharepoint.com/:u:/g/personal/karoliij_ut_ee/EQVIO7ltiXJIg3RAO1pp1dYB32XHSlIk4dhH43QCmyeE-Q?e=D7poe5). Seda saab importida INCEpTIONi versiooni 35.0 või uuemasse versiooni, lisaks on kaustas olemas tsv-failid (iga faili kaustas on sobiv fail nimega „admin.tsv“).

Programmeerimiskeelega R (R Core Team 2020) tegin ma tabeli, kus oli iga püsiühendiga seonduvad andmed, ning programmeerimiskeelega Python (Python Software Foundation s.a.) lisasin uutesse failidesse juba leitud püsiühendid ja juba märgendatud failidesse kahe silma

[4] AdjID = omadussõnafraasiline, AdvID = määrsõnafraasiline, ModMWE.VID = deverbaalne modifitseerijafraasiline, NID = nimisõnafraasiline, PronID = asesõnaline, VMWENom = deverbaalne nimisõnafraasiline. Lühendid on võetud UniDive’i juhendist. (Savary jt *s. a.* (a): 8–10)

vahele jäänud püsiühendid. Andmete puhastamisel lisasin muuhulgas ainult osaliselt püsiühendiks märgitud liitsõnadele, nt *maailm* liitsõnas *muusikamaailm*, püsiühendireale õige lemma ning koondasin mitmesõnaliste püsiühendite andmed ja lemmad ühele reale. R-is tegin ka joonised pakettidega ggplot2 (Wickham 2016) ja gridExtra (Auguie 2017), leidsin sagedused ja tegin tabeli iga unikaalse püsiühendi lemma vormide, (eri tekstiliikides) esinemissageduste, liigi, testi ja testi näitega, millest vähemalt kaks korda esinevad panin Lisasse 3.

5.1. Märgendamisel esile kerkinud probleemkohad

Üks küsimus, mis ette tuli, oli produktiivsuse piiritlemine ehk kui produktiivne peaks ühendi mall olema, et selle variante ei loetaks mitte eraldi püsiühenditeks vaid tavalisteks konstruktsioonideks. Kui variante on alla kümne, on lugu selge, aga rohkemate puhul aitasid otsustada moodustamise (semantilised) tingimused ja kui paljud variantidest on tegelikult sagedased. Näiteks *aeg-ajalt* ja *aasta-aastalt* laadi ühendites peab korduv sõna olema järjestatav või korduv ja ainult 7 sellist ühendit esines eesti keele ühendkorpuses 2023 (Koppel jt 2023) üle tuhande korra (esinemissageduste leidmiseks kasutasin Sketch Engine'i (Kilgarriff jt 2014)). Samas *potitäis* ja *karbitäis* laadi liitsõnades peab esiosa olema mahuti ja 42 sellist ühendit esines eesti keele ühendkorpuses üle tuhande korra. Seetõttu lugesin *aegajalt* ja *aeg-ajalt* püsiühenditeks, aga näiteks *suu-* ja *lusikatäit* mitte. Sarnaselt peab *vastu tänast* ja *vastu hommikut* laadi sõnaühendites põhiosa olema lihtsalt ajasõna ja 11 sellist ühendit esines ühendkorpuses üle tuhande korra, nii et seda püsiühendiks ei lugenud.

Teine küsimus, mis tekkis, oli nimisõnafraasiliste püsiühendite kohta, mis on tuletatud omadussõnadest, mis on omakorda tuletatud tegusõnadest. Näiteks *ettevaatlikkus* on tuletatud omadusõnast *ettevaatlik*, mis on tuletatud tegusõnast *ette vaatama*, ning *tõsiseltvõetavus* on tuletatud omadussõnast *tõsiseltvõetav*, mis on omakorda tuletatud tegusõnast *tõsiselt võtma*. Kuna nende aluseks on siiski verbifraasid, siis olen neid lugenud deverbaalseteks nimisõnalisteks, mitte lihtsalt nimisõnalisteks püsiühenditeks.

Kolmandaks probleemkohaks osutus erialasõnavara, millega tekkis mitmeid probleeme. Esiteks, mida lugeda liitsõna mõne komponendi algseks tähenduseks, näiteks *kava* sõnas *mõõtkava* on ilmselt tähenduses 'plaan' ja *raadio* sõnas *raadiogalaktika* on ilmselt lühendatud *raadiokiirgusest*. Esimese puhul saab järelikult teha sõnaasendustesti, aga teise puhul ei leidnud ma sobivat sõna, millega seda asendada või täiendada, seega ei saanud see üheltki testilt positiivset tulemust ehk tegu pole püsiühendiga. Ühel korral oli aga ühendi ühte

komponenti eelmises lauses selles kontekstis defineeritud, mistõttu oli seda kergem mitte-püsiühendiks määrata, nimelt *kõrv* ühendis *kahekõrvaline galaktika* defineeriti kui ’galaktika keskmest välja paiskunud võimas gaasijuga või -pilv’. Teine probleem oli see, kui terminit on väljaspool konkreetset teksti nii vähe mainitud, nii et selle täpsest tähendusest on raske aru saada, mõned sellised olid näiteks *hõimuusuline* ja *iduliin*. Kolmas probleem tekkis *valgusaasta* puhul, millel on ainult statistiline eripära ehk kuigi pole termineid *valgussajand* ega *valguskuu*, siis ei oleks need ebagrammatilised ega tekitaks tähendusnihet, lihtsalt on otsustatud, et valgust mõõdetakse aastates, mitte sajandites ega kuudes.

Neljas küsimuskoht tekkis [nimi] + [liigisõna] ühendite juures, nagu *Parkinsoni tõbi* ja *Nobeli preemia*, ehk kas neid lugeda nimeüksusteks või püsiühenditeks. Kui vaadata nimeüksuste teste, siis esimene test on selle kohta, kas antud kandidaadid viitavad selles kontekstis spetsiifilistele entiteetidele. *Parkinsoni tõbi* kontekstis *tahame näiteks Parkinsoni tõbe ravida* ilmselt jah, *Nobeli preemia* kontekstis *võiksin Nobeli preemiale kandideerida* samuti ilmselt jah. Seega peaks edasi liikuma testi NMWE.4 juurde, mille puhul tundub mõlemal ühendil vastuseks olevat jah, sest mitmetel inimestel on *Parkinsoni tõbi* või *Nobeli preemia*. Järgmises testis, NMWE.5, küsitakse konkreetselt, kas tegu on isiku, organisatsiooni, asukoha, inimtoote või sündmusega. *Parkinsoni tõbi* ei ole ükski neist, aga *Nobeli preemia* puhul on vastamine keerulisem, sest seda võib ehk lugeda mitte-füüsiliseks inimtooteks. Seega *Parkinsoni tõbe* võib edasi käsitleda püsiühendikandidaadina, aga *Nobeli preemiat* mitte.

Viiendaks probleemkohaks olid *pea*-liitega nimisõnad, nagu *pealinn, peatoimetaja* ja *peaminister*. Liitsõna *pealinn* on otsetõlge saksakeelsest liitsõnast *Haupstadt* (Sõnaveeb 2025a), kus *Haup* tähendab pead ja *Stadt* linna, seega see on tõlgitud püsiühend. Teised sarnased *pea*-liitega liitsõnad on ilmselt tekkinud *pealinna* eeskujul, kuigi nende päritolu pole Sõnaveebis kirjas. Samas tõlgendatakse *pea*-liidet tänapäeval pigem kui lühendatud vormi sõnast *peamine* (Sõnaveeb 2025b) ja seda ka *pealinna* puhul (Kasik 2015: 246). Seega olen lähtunud tänapäevasest tõlgendusest, mille järgi pole need liitsõnad püsiühendid, ja käsitlenud segaduse vältimiseks ka *pealinna* samamoodi.

Kuues küsimus tekkis selliste sarnaste ühendite kohta nagu *suurlinn, suuromanik, suur laul* ja *suur kunstnik*. Otsustasin lugeda liitsõnad püsiühenditeks ja mitmesõnalised ühendid mitte, sest *suurlinn* ja (*väga*) *suur linn* puhul on sõnal *suur* erinev tähendus, aga *suur kunstnik* ja *väga suur kunstnik* puhul seda tähenduseristust ei teki.

Lisaks oli ka üksikuid juhtumeid, mis võisid olla mitmetimõistetavad ja mille puhul pidin lähenemist täpsustama. Näiteks võistlustel auhinnalisi kohti käsitlesin vastavalt sellele, kas viidati medalitele, näiteks *pronksi [pronks medali] saanud laevastik*, või ei, näiteks

pronkslaevastik, sest medal on tõesti pronksist, aga laevastik pole. *Jahimees*-tüüpi liitsõnades käsitlesin põhisõna *mees* tähendust kui ’inimene’, mitte ’meessoost inimene’, ja seetõttu said neist mitte-püsiühendid. Liitsõnas *asetäitja* käsitlesin *ase* tähendust kui ’aseme (täitja)’, mitte ’asendav (täitja)’, mida see liitsõna esiosana tavaliselt tähendab, ja seega ei jõudnud ka see liitsõna püsiühendite sekka.

6. Tulemuste analüüs

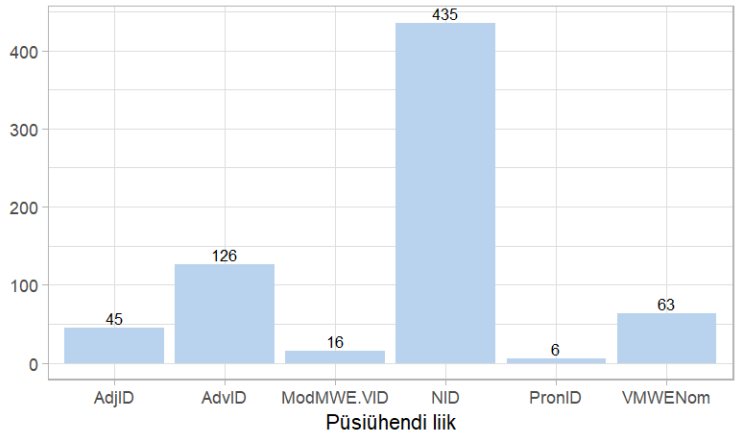
Märgendasin 18 tuhande sõnelises ilukirjandus-, ajakirjandus- ja populaarteaduslikest tekstidest koosnevas korpuses nimi-, omadus, ja määrsõnafraasina kasutatavaid püsiühendeid, mille alla kuulusid ka asesõnalised ja deverbaalsed püsiühendid. Leidsin 691 püsiühendit, mille seas oli 472 unikaalset püsiühendivormi ja 356 unikaalset püsiühendilemmat. Järgnevalt analüüsin tulemusi uurimisküsimustest lähtuvalt: peatükis 5.1 toon välja, millised püsiühendid ja mis liiki püsiühendid korpuses kõige rohkem esinesid; peatükis 5.1 analüüsin, kuidas püsiühendite kasutus registrites erines; peatükis 5.2 vaatlen, millised testid kõige rohkem püsiühendeid tuvastasid, ja peatükis 5.4 uurin liitsõnade ja püsiühendite vahekorda, sest käsitlesin ka liitsõnu võimalike püsiühenditena.

Kuigi lauseid oli igast tekstiliigist sama palju, siis sõnede arv varieerus umbes viiesaja sõne ulatuses, seega võrdlesin püsiühendite esinemissagedusi normaliseeritud esinemissagedustega tekstiliigiti. Normaliseerimine tähendab, et kui korpused on erineva suurusega, siis nende võrdlemiseks vaadatakse, kui palju esineks uuritavat nähtust sama suures korpuses ehk baasis (Muischnek & Lindström 2020: 331–332). Võtsin baasiks kuus tuhat sõnet, seega arvutasin, kui palju esineks püsiühendeid 5600-sõnelise korpuse asemel 6000-sõnelises korpuses. Kuna püsiühendite absoluutsagedused ja normaliseeritud sagedused erinesid üsna vähe (alla 7%, vt Tabel 1), siis otsustasin analüüsida siiski absoluutsagedusi.

Tabel 1. Püsiühendite sagedus korpuses ja normaliseeritud sagedus tekstiliigiti.

	Ajakirjandus	**Ilukirjandus**	**Populaarteadus**
Sagedus korpuses	247	181	263
Normaliseeritud sagedus	240	193	248
Erinevus	-3,0%	6,7%	-5,7%

6.1. Püsiühendite sagedused ja liigid



Joonis 1. Püsiühendite liikide[5] sagedused.

Püsiühendid jagunesid testimise tulemusel eelnevalt seletatud meetodite alusel kuude kategooriasse: nimisõnalised, asesõnalised, deverbaalsed nimisõnalised, omadussõnalised, määrsõnalised ja deverbaalsed modifitseerijafraasilised. Joonisel 1 on välja toodud eri liiki püsiühendite esinemissagedused. Korpuses esines kaugelt kõige rohkem nimisõnafraasilisi püsiühendeid, mida esines üle neljasaja korra. Üle saja korra esines määrsõnafraasilisi ja üle viiekümne korra deverbaalseid nimisõnafraasilisi püsiühendeid, üle neljakümne korra omadussõnafraasilisi, kuusteist korda deverbaalseid määr- või omadussõnafraasilisi ja alla kümne korra asesõnalisi püsiühendeid. Deverbaalsed modifitseerijafraasilised püsiühendid olid ülekaalukalt omadussõnafraasilised, ainult kaks olid määrsõnafraasilised.

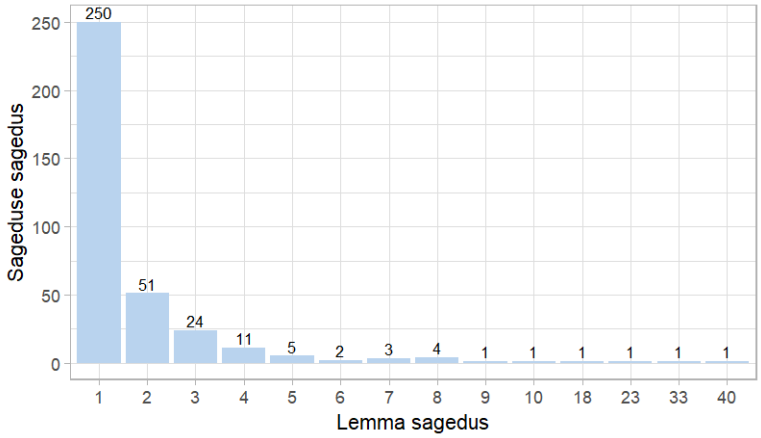
Kõige sagedasemad nimisõnafraasilised püsiühendid olid *maailm* (esines 37 korda), *must auk* (33) ja *piirikoer* (23), viimased kaks esinesid konkreetsetes tekstides, mis olid just nende kohta. Sagedasemad deverbaalsed nimisõnafraasilised püsiühendid olid *kokkulepe* (8), *tähelepanu*, *üleminek* ja *piiririkkuja*, viimased kolm esinesid kolm korda, ning määrsõnafraasilised *peaaegu* (18), *eelkõige* (8), *esialgu* (6) ja *otsekui* (5), mitmesõnalistest esines üle ühe korra ainult *kogu aeg* (2). Omadussõnafraasilistest püsiühenditest aga ainult kuus esineski rohkem kui ühe korra, sagedasim oli *ajalooline*, mis esines kolm korda.

Deverbaalsete modifitseerijafraasiliste püsiühendite seas esines ainult üks korduv tüvi, *kättesaadav*, mis esines lemmadena *kättesaadav* ja *kättesaadavam*, asesõnalistest esines aga *igaüks* lemmana neli korda, lisaks esines korra *iseenese* ja *teineteise*. Üle ühe korra esines

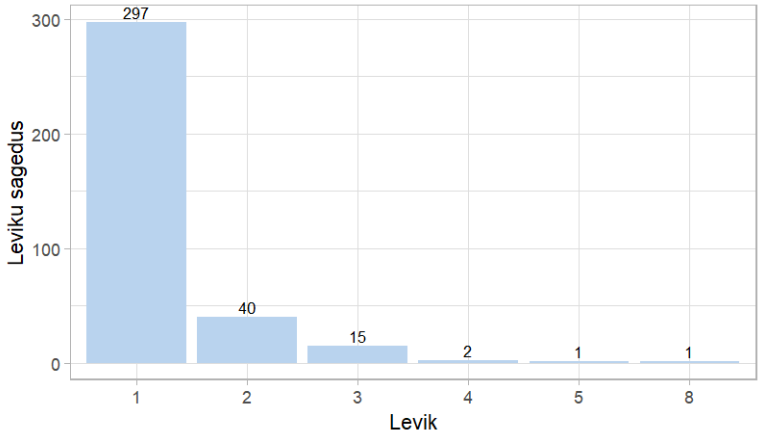
[5] AdjID = omadussõnafraasiline, AdvID = määrsõnafraasiline, ModMWE.VID = deverbaalne modifitseerijafraasiline, NID = nimisõnafraasiline, PronID = asesõnaline, VMWENom = deverbaalne nimisõnafraasiline. Lühendid on võetud UniDive'i juhendist. (Savary jt *s. a.* (a): 8–10)

14,9% (65 lemmat) nimisõnafraasilistest, 17,5% (11 lemmat) deverbaalsetest nimisõnafraasilistest, 13,3% (6 lemmat) omadussõnafraasilistest ja 17,5% (22 lemmat) määrsõnafraasilistest püsiühenditest. Samas oli paar juhtu, kus sama sõna esines nii sidekriipsuga kui ka ilma, nimelt *võib-olla* (1 kord) ja *võibolla* (3 korda) ning *aeg(-)ajalt* (mõlemat 1 kord).

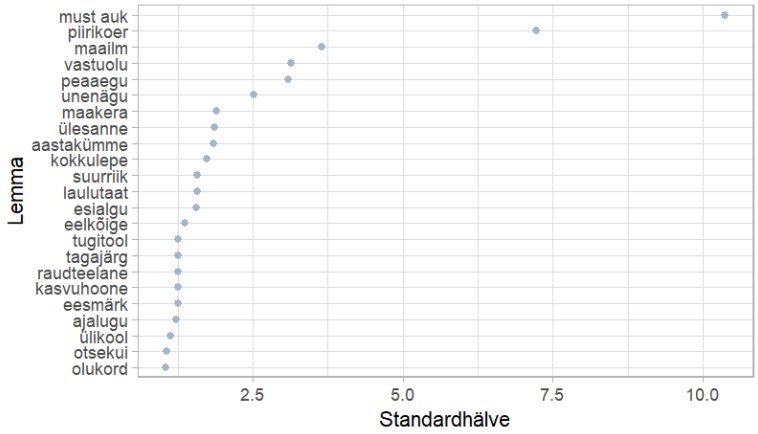
Kui vaadata kõiki püsiühendite liike koos, nagu joonisel 2, siis umbes 70% püsiühendite lemmadest esineb ainult ühe korra ja ainult viis esineb vähemalt kümme korda, millest neli on nimisõnalised. Joonis 3 illustreerib lemmade leviku jaotumist ehk seda, mitmes üheksast failis lemmat esines (Brezina 2018: 48). Sellelt on näha, et veel suurem osakaal lemmadest, lausa 83,4%, esines ainult ühes failis ja üle pooltes failides ehk vähemalt neljas failis esines vaid neli lemmat, milleks olid leviku suuruse järjekorras *maailm*, *olukord*, *ajalugu* ja *igaüks*. Ka kolmes failis esinenutest ainult üks oli omadussõnafraasiline (*ajalooline*) ja kuus määrsõnafraasilised, näiteks *peaaegu* ja *tõenäoliselt*.



Joonis 2. Lemmade sageduste sagedused.



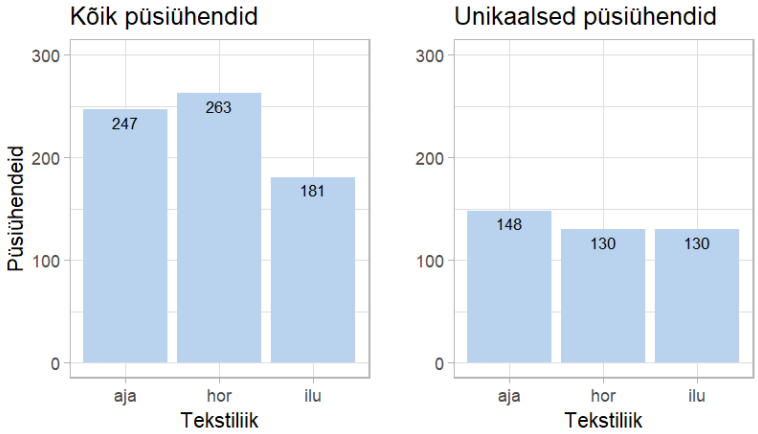
Joonis 3. Levikute sagedused.



Joonis 4. Suurima standardhälbega püsiühendilemmad.

Samas toob joonis 4 välja kõige suurema standardhälbega püsiühendilemmad. Standardhälve näitab, kui palju erineb lemma esinemissagedus keskmiselt eri failides võrreldes selle keskmise esinemissagedusega (Brezina 2018: 48–49). Näiteks esinevad *must auk* ja *piirikoer* ainult ühes failis hästi palju, mistõttu on nende standardhälve suur. Samas *maailm* esineb küll pea kõigis failides, aga selle sagedus neis varieerub ühe ja üheteistkümne vahel, ning *peaaegu*, suurima standardhälbega määrsõnafraasiline püsiühend, esineb kolmes failis ja selle sagedus neis varieerub viie võrra.

6.2. Püsiühendid eri tekstiliikides



Joonis 5. Kõik ja unikaalsed püsiühendid eri tekstiliikides[6].

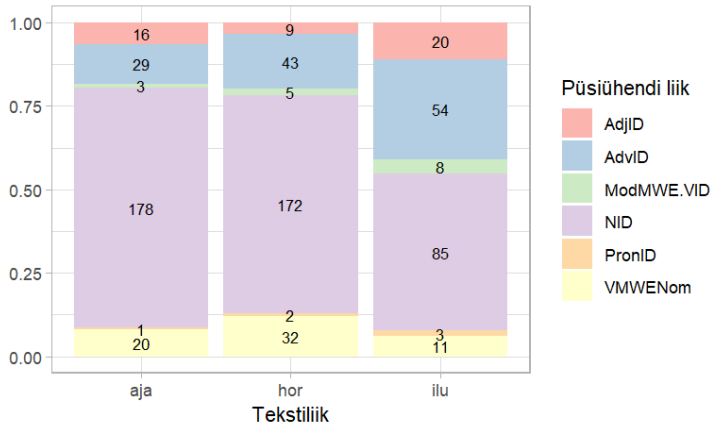
Kuna märgendasin püsiühendeid kolmes tekstiliigis, ajakirjanduses, ilukirjanduses ja populaarteaduses, siis ongi joonisel 5 välja toodud kõikide püsiühendite esinemised tekstiliigiti

[6] Aja = ajakirjanduslikud tekstid, hor = populaarteaduslikud tekstid, ilu = ilukirjanduslikud tekstid. Lühendid on võetud eesti keele UD puudepangast, kust tekstid pärit on. (UD *s. a.*)

ja unikaalsete lemmade hulk tekstiliigiti. Kuigi kõige rohkem püsiühendeid esines populaarteaduslikes tekstides, siis ajakirjanduslikes tekstides varieerusid need rohkem. Selle põhjuseks võib olla püsiva terminoloogia kasutamine, näiteks ühes populaarteaduslikus artiklis räägiti palju mustadest aukudest ja kasutatigi kogu aeg seda sõna, aga ajaleheartiklis, kus räägiti piirikoertest, kasutati ka näiteks sõnu *teenistuskoer* ja *piirivalvekoer*. Huvitav on ka see, et ilukirjanduslikes tekstides kasutati püsiühendeid vähem, kuigi olemasolevad varieerusid sama palju kui populaarteaduslike tekstide omad.

Kui mõelda võimalike põhjuste peale, siis üheks võib olla valitud autorite stiilid, kuigi ilmselt mitte nende vähene hulk, sest horisondi tekstidel olika umbes neli autorit (paar lõiku ka viiendalt). Teine võimalik põhjus võib olla rohkemate metafooride ja võrdluste kasutamine ilukirjanduses, näiteks *siis võid koondamishüvitist näha nagu oma kõrvu*, või pikemad kirjeldused, mis kasutavad lihtsamaid sõnu, näiteks *tundes tolle halli kivi kõrvetavat külmust*. Samas on selliste lauseüksuste sagedusi keeruline kiiresti leida ja võrrelda.

Ajakirjanduslikes tekstides olid kõige sagedasemad lemmad *piirikoer* (esines 23 korda), *maailm* (6), *ajalugu* (6) ja *ülesanne* (6), ilukirjanduslikes tekstides *peaaegu* (9), *unenägu* (8), *maailm* (5) ja *laulutaat* (5) ning populaarteaduslikes tekstides *must auk* (33), *maailm* (19), *vastuolu* (10) ja *aastakümme* (7). Nagu näha, siis kõige sagedasem sõna oli igas tekstiliigis erinev, aga oli ka üldiselt sagedasi sõnu, nagu *maailm* ja *peaaegu*, millest esimene esines kõigis tekstiliikides üle viie ja teine üle nelja korra.



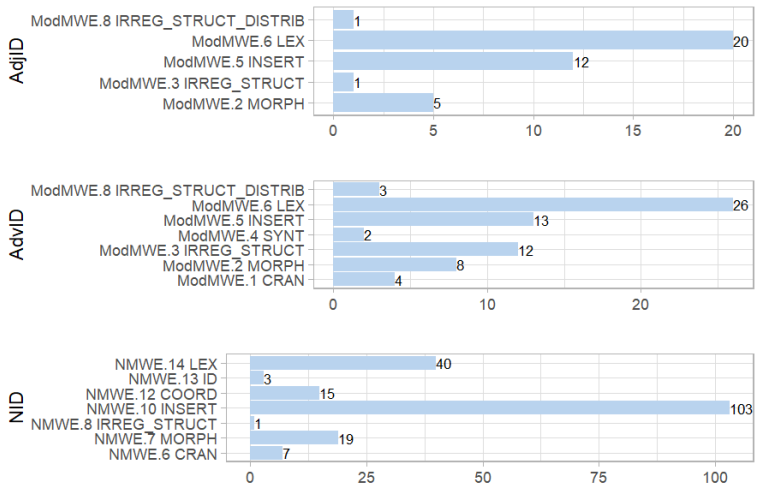
Joonis 6. Püsiühendite liikide[7] jaotus tekstiliigiti[8].

[7] AdjID = omadussõnafraasiline, AdvID = määrsõnafraasiline, ModMWE.VID = deverbaalne modifitseerijafraasiline, NID = nimisõnafraasiline, PronID = asesõnaline, VMWENom = deverbaalne nimisõnafraasiline. Lühendid on võetud UniDive'i juhendist. (Savary jt *s. a.* (a): 8–10)

[8] Aja = ajakirjanduslikud tekstid, hor = populaarteaduslikud tekstid, ilu = ilukirjanduslikud tekstid. Lühendid on võetud eesti keele UD puudepangast, kust tekstid pärit on. (UD *s. a.*)

Joonis 6 illustreerib püsiühendite liikide osakaale eri tekstiliikides, tuues välja ka esinemiste absoluutsagedused. Kuigi kõigis tekstiliikides esineb nimisõnafraasilisi püsiühendeid kõige rohkem, siis esineb populaarteaduslikes tekstides rohkem deverbaalseid nimisõnafraasilisi püsiühendeid, nagu *kokkulepe* ja *tähelepanu*, ning vähem omadussõnafraasilisi püsiühendeid, nagu *ajalooline* või *otstarbekas*. Samuti on huvitav, et määrsõnafraasilisi püsiühendeid, nagu *omaette*, *otsekui* ja *üksisilmi*, esineb ajakirjanduslikes tekstides võrdlemisi vähem ja ilukirjanduslikes tekstides võrdlemisi rohkem.

6.3. Püsiühendite testid



Joonis 7. Unikaalseid püsiühendeid tuvastanud testide sagedused.

Kui peatükis 4 kirjeldasin UniDive'i teste ja kuidas neid oma töö jaoks kohandasin, tuues iga testi tuvastatud püsiühendite kohta ka paar näidet, siis joonis 7 illustreerib, kui palju unikaalseid lemmasid eri testid tuvastasid. Erinevaid nimisõnafraasilisi püsiühendeid tuvastas kõige rohkem laiendi lisamise test, seejärel sõna asenduse test, vormi muutmise test ja koordinatsiooni test. Siin tuleb küll arvestada, et kuna teste tuli läbida kindlas järjekorras, siis lõpetati püsiühendiga tegelemine pärast mõnelt testilt positiivse vastuse saamist, v.a ebatavaliste sisemiste lauseliikmete test ModMWE.8 IRREG_STRUCT_DISTRIB (vt põhjendust peatükis 4.2). Leitud nimisõnafraasilistest püsiühenditest ühtegi ei tuvastanud süntaktilise muutmise test NMWE.11 SYNT ega ebatavaliste sisemiste lauseliikmete test NMWE.9 IRREG_STRUCT_DISTRIB, kuigi mõlemale on võimalik eesti keeles näiteid leida, näiteks esimesele *maa ja ilm*, sest ei saa öelda *ilm ja maa*, ning teisele *meelespea*, millel on sisemine verbifraasi struktuur, kuigi süntaktiliselt käitub nimisõnana.

Omadus- ja määrsõnafraasilisi püsiühendeid tuvastasid kõige rohkem sõna asenduse test ja laiendi lisamise test. Sõna asenduse testi esines nende puhul ilmselt rohkem kui laiendi lisamise testi sellepärast, et muutumatutele sõnadele on raske laiendeid leida. Näiteks püsiühendites *niivõrd* ja *alatihti* on raske leida laiendeid sõnadele *nii* ja *ala*, aga saab leida teoreetiliselt võimalikud sõnaasendused, nagu *nõndavõrd* ja *alasageli*, mida siiski ei eksisteeri. Mitmeid omadussõnafraasilisi püsiühendeid tuvastas ka vormi muutmise test ja määrsõnafraasilisi ebatavalise struktuuri test. Leitud omadussõnafraasilistest püsiühenditest ühtegi ei tuvastanud süntaktilise muutmise test ega jäänukmorfi test, neile ei ole ma ka raske eesti keelest näiteid leidnud.

Kõige sagedamini on püsiühendi tunnustest tuvastatav leksikaalne, süntaktiline ja vormiline idiosünkraatilisus ning sellest tulenev paindumatus. Püsiühendid on enamasti kinnistunud osade või vormiga ja selle osad ei käitu süntaktiliselt produktiivselt, mis tuleb välja nii täiendite lisamisest tuleneva tähendusliku nihke või võimatuse näol kui ka näiteks ühendi sisemise struktuuri ebatavalisuse näol. Leksikaalse idiosünkraatilisuse näiteks on *kõrvulukustav* ja olematu liitsõna *silmilukustav*, vormilise näiteks *tänapäev* ja selle (tähenduslikult) võimatu vorm *tänapäevad*, süntaktilise näiteks *must auk* ja *üleni must auk* ning sisemise struktuuri näiteks *seepärast* ja *sellepärast*. Mõne puhul pole teist varianti keeles olemas, osade puhul on see aga muu tähendusega või illustreerib esimese variandi ebatavalisust.

6.4. Liitsõnalisus

Minnes edasi püsiühendite vormistuse juurde, siis kui arvestada iga esinemiskorda eraldi, olid leitud püsiühenditest 89,3% liitsõnad ja ülejäänud mitmesõnalised. Kui arvestada iga lemmat ainult üks kord, siis olid leitud püsiühenditest 86,2% liitsõnad. Sarnaselt, kui arvestada iga esinemiskorda eraldi, siis olid esinenud liitsõnadest 30,7% püsiühendid, ning kui arvestada igat lemmat ainult üks kord, siis olid esinenud liitsõnade lemmadest 14,2% püsiühendid. Ülejäänud olid mitte-püsiühendilised liitsõnad, nagu *teatriomanik* või *veealune*. Liitsõnade hulka on arvatud ka sidekriipsuga ühendatud sõnad, sest on ühendeid, millel on kokkukirjutatud ja sidekriipsuga rööpvormid, nagu *võib-olla* ja *võibolla*. Lisaks oli sidekriipsuga püsiühendeid ainult seitse ja sidekriipsuga sõnu 43, kui välja jätta nimed ja numbrit sisaldavad sõnad.

Juba märgendamise jooksul oli aru saada, et suur osa püsiühenditest on liitsõnad, aga see, kui paljud liitsõnad olid püsiühendid, oli minu jaoks üllatav. Samas tuleb arvestada ka seda, et kuna eesti keeles on (liit)sõnamoodustus üsna produktiivne (Kull 1967: 1; Kasik 2015: 28), siis

on püsivatel nimisõnaühenditel kalduvus liitsõnastuda (Kasik 2015: 56). Seda võib juhtuda näiteks siis, kui ühendi esiosa on genitiivis, kuna siis see osa eraldi ei käändu. Võrrelda võib näiteks ühendeid *must auk* ja *piirikoer*, mis adessiivis on *mustal augul* ja *piirikoeral*. Lisaks võib genitiivis esiosaga ühendi osade lahku kirjutamine viidata hoopis omamissuhtele, näiteks *lasteaed* ja *laste aed*.

Tekstiliikidest oli kõige vähem mitte-liitsõnalisi püsiühendeid ajakirjanduslikes ja kõige rohkem populaarteaduslikes tekstides, viimase puhul *musta augu* sageduse tõttu. Unikaalsetest lemmadest oli kõigis tekstiliikides üsna sama osakaaluga liitsõnu. Kõik asesõnalised püsiühendid olid liitsõnad ja kõige suurem mitmesõnaliste osakaal oli omadus- ja määrsõnafraasilistel püsiühenditel. Mõned näited neist on *kullastki kallim, surmani väsinud, kas või* ja *omal ajal*.

Testide puhul ei olnud ühtegi liitsõnalist püsiühendit süntaktilise muutmise testi tuvastatud püsiühendite hulgas, milleks olid *ikka ja jälle* ning *risti ja põiki*. Lisaks olid enamik modifitseerijafraasilised vormi muutmise testiga tuvastatud püsiühendid mitmesõnalised, sest sellised sisaldavad sagedamini muudetavaid sõnu. Kõik ebatavalise struktuuri testiga ja koordinatsioonitestiga leitud nimisõnalised ning jäänukmorfi testiga leitud modifitseerijafraasilised püsiühendid olid liitsõnalised. Põhjuseks on ilmselt see, et koordinatsiooni testi saabki teha ainult liitsõnalistele ja teisi esines alla viie.

Kokkuvõte

Töö eesmärk oli teada saada, millised on eestikeelsetes tekstides sagedasemate püsiühendite sõnaliigid ja kuju (liitsõna, mida käsitlesin selles töös sõnaühendina, või lahku kirjutatud sõnade ühend) ning kuidas need erinevad tekstiliigiti. Lisaeesmärk oli teada saada, kuidas on rahvusvahelise UniDive'i initsiatiivi püsiühendite tuvastamise testid kohandatavad eesti keele püsiühendite tuvastamiseks. Uurimisküsimused olid järgmised:

1. Millised püsiühendid esinevad tekstis kõige rohkem, sealhulgas sõnaliigi poolest?
2. Kuidas erineb püsiühendite kasutus erinevates registrites (ajalehetekstid, ilukirjandus- ja populaarteaduslikud tekstid)?
3. Kuidas saab UniDive'i initsiatiivi püsiühenditeste eesti keele jaoks kohandada ja millised testid tuvastavad kõige rohkem püsiühendeid?
4. Kui paljud leitavatest püsiühenditest on liitsõnad ja kui paljud esinevatest liitsõnadest on püsiühendid?

Selleks märgendasin eesti keele Universal Dependencies puudepanga (Zeman jt 2024) igas valitud tekstiliigis (ilukirjandus-, ajakirjandus- ja populaarteaduslikes tekstides) viiesajas lauses nimi-, omadus- ja määrsõnafraasilisi püsiühendeid. Nende leidmiseks kasutasin ja kohandasin UniDive'i initsiatiivist 15 testi nimisõnafraasiliste leidmiseks ja 8 testi modifitseerijafraasiliste leidmiseks (Savary jt *s. a.* (b): 3–8). Teste rakendati kindlas järjekorras, kuni mõni sai positiivse tulemuse. Märgendatud püsiühendid jagunesid lausedistributsiooni järgi nimisõnafraasilisteks, mille alamliigid olid nimisõnalised, asesõnalised ja deverbaalsed nimisõnalised, ning modifitseejafraasilisteks, alamliikidega omadussõnalised, määrsõnalised ja deverbaalsed modifitseerijad.

Minu töö tulemuseks on eesti keele jaoks kohandatud püsiühendite testid, nende testide alusel märgendatud korpus ning teadmised püsiühendite esinemissagedusest, sh püsiühendite liikide ja tekstiliikide kaupa. Nimisõnafraasiliste püsiühenditestide juures tuli kohandada asesõnaliste püsiühendite leidmise testi, kuna see põhines eelnevalt olemasoleval asesõnaliste püsiühendite nimekirjal, mida eesti keeles olemas pole. Seetõttu käisin asesõnaliste püsiühendikandidaatidega ka ülejäänud testid läbi, kuni mõni kandidaadi püsiühendiks määras, aga testiks märkisin ikkagi asesõnalisuse testi. Edaspidi oleks mõistlik teha eestikeelsete asesõnaliste püsiühendite list või käsitleda nimi- ja asesõnalisi püsiühendeid sarnaselt modifitseerijafraasilistele, kus püsiühendi täpsem liik määratakse peale mõnelt testilt positiivse tulemuse saamist.

Modifitseerijafraasiliste püsiühenditestide juures lisasin märgendamise jooksul vastavalt vajadusele paar testi, mis olid nimisõnafraasiliste püsiühenditestide juures, kuid sobivate püsiühendite väikese esinemistõenäosuse tõttu modifitseerijafraasiliste püsiühenditestide nimekirjast välja jäetud. Testide lisamise tõttu muutsin veidi testimise järjekorda. Tulevikus oleks mõistlik muuta testide järjekorda loogilisemaks.

Leidsin 18 tuhande sõnelisest korpusest 691 püsiühendit, milles oli 472 unikaalset püsiühendivormi ja 356 unikaalset püsiühendilemmat. Ligi 90% leitud püsiühenditest olid liitsõnad ja esinenud liitsõnadest 31% olid püsiühendid, kuigi unikaalsete lemmade järgi vaadates langeb osakaal 14%-le. Veidi rohkem liitsõnalisi püsiühendeid oli ajakirjanduslikes tekstides. Samuti olid kõik asesõnalised püsiühendid ning kolme testiga tuvastatud püsiühendid liitsõnad. Vähem liitsõnu oli sageduselt populaarteaduslikes tekstides. Kõik süntaktilise muutmise testiga tuvastatud püsiühendid olid lahku kirjutatud.

Üle poole leitud püsiühenditest olid nimisõnalised, alla viiendiku oli määrsõnalisi, ülejäänuid oli alla kümnendiku. Kõige sagedasem nimisõnaline püsiühend (maailm) esines 37 korda, määrsõnaline (*peaaegu*) 18 korda, deverbaalne nimisõnaline (*kokkulepe*) kaheksa korda, omadussõnaline (*ajalooline*) kolm korda, asesõnaline (*igaüks*) kolm korda ja deverbaalne modifitseerijafraasiline kaks korda. Viimase puhul oli tegu sama tüve eri kujudega – *kättesaadav* ja *kättesaadavam*. Samas esines 70% püsiühendilemmadest vaid korra ja üle 80% esines vaid ühes failis üheksast, nende hulgas kaks kolmest kõige sagedasemast püsiühendist.

Tekstiliigiti esines kõige rohkem püsiühendeid populaarteaduslikes tekstides ja kõige vähem ilukirjanduslikes tekstides. Unikaalseid lemmasid oli kõige rohkem ajakirjanduslikes tekstides, kusjuures populaarteaduslikes ja ilukirjandustekstides oli neid sama palju. Populaarteaduslikes tekstides võis olla rohkem korduvaid lemmasid seetõttu, et seal kasutati terminoloogiat püsivamalt kui ajakirjanduslikes tekstides, kus varieeriti sõnakasutust rohkem. Lisaks oli ilukirjandustekstides suurem modifitseerijafraasiliste ja väiksem nimisõnafraasiliste püsiühendite osakaal kui teistes tekstiliikides.

Nimisõnafraasilisi püsiühendeid tuvastas kõige rohkem täiendi lisamise test, seejärel sõna asendamise test ja vormi muutmise test. Modifitseerijafraasilisi tuvastas kõige rohkem sõna asendamise test ja alles siis täiendi lisamise test, ilmselt sellepärast, et muutumatutele sõnadele on raske laiendeid leida. Ühtegi leitud nimisõnafraasilistest püsiühenditest ei leidnud süntaktilise muutmise test ega ka ebatavaliste sisemiste lauseliikmete test, kuigi neile on eesti keeles näiteid olemas. Ühtegi omadussõnafraasilist püsiühendit ei leidnud süntaktilise muutmise test ega ka jäänukmorfi test, millele ma eestikeelseid näiteid ei leidnud.

Tulevikus oleks huvitav samu asju uurida suurema koguse materjali abil, et näha, kas ja kuidas tulemused muutuvad. Lisaks võiks kaasata uurimisse teisi püsiühendiliike, nagu verbi- ja funktsioonifraasilisi (näiteks sidesõna- või hüüdsõnafraasilisi), ning muuta testide järjekorda. Samuti oleks huvitav võrrelda tulemusi teiste keelte samal raamistikul põhinevate tulemustega.

Kirjandus

Aedmaa, Eleri. 2019. *Detecting compositionality of Estonian particle verbs with statistical and linguistic methods* (Dissertationes Linguisticae Universitatis Tartuensis 37). Tartu: University of Tartu Press.

Auguie, Babtiste. 2017. *gridExtra: Miscellaneous Functions for „Grid“ Graphics*. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. 1. tr. Cambridge University Press. https://doi.org/10.1017/9781316410899.

COST. *s. a.* Action CA21167. https://www.cost.eu/actions/CA21167. (Vaadatud 21.02.2025).

De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 1–54. https://doi.org/10.1162/coli_a_00402.

Dovgan, Lina. 2013. *Fraseologismide tõlkimismustrid ungari-eesti ja ungari-ukraina sõnaraamatu materjali põhjal*. Tartu Ülikool, eesti ja üldkeeleteaduse instituut. Magistritöö. http://hdl.handle.net/10062/33782.

Erelt, Mati, Tiiu Erelt & Kristiina Ross. 2020. *Eesti keele käsiraamat*. Uuendatud väljaanne. Tallinn: Eesti Keele Instituut.

Granger, Sylviane & Fanny Meunire (toim). 2008. *Phraseology: an interdisciplinary perspective*. *z.139*. John Benjamins Publishing Company. https://benjamins.com/catalog/z.139. (Vaadatud 29.10.2024).

Gries, Stefan Th. 2008. Phraseology and linguistic theory: A brief survey. Fanny Meunire & Sylviane Granger (toim), *Phraseology: an interdisciplinary perspective*, 3–25. John Benjamins Publishing Company. https://benjamins.com/catalog/z.139. (Vaadatud 29.10.2024).

Kasik, Reet. 2015. *Sõnamoodustus* (Eesti keele varamu 1). Tartu: Tartu Ülikool Kirjastus.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář & Vít Suchomel. 2014. The Sketch Engine: Ten Years On. http://www.sketchengine.eu.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho & Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation.

Koppel, Kristina, Jelena Kallas, Madis Jürviste & Helen Kaljumäe. 2023. Eesti keele ühendkorpus 2023. Lexical Computing Ltd. / Eesti Keele Instituut.

Kull, Rein. 1967. *Liitsõnade kujunemine eesti kirjakeeles*. Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Lexicalized components and open slots. *s. a.* Lexicalized components and open slots. *PARSEME Shared Task 1.2 - Annotation guidelines*. https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=lexicalized#lexicalized. (Vaadatud 15.11.2024).

Mel’čuk, Igor. 2001. Collocations and Lexical Functions. Anthony Paul Cowie (toim), *Phraseology: theory, analysis, and applications* (Oxford Studies in Lexicography and Lexicology), 23–53. paperback ed. Oxford: Oxford Univ. Press.

Mensalo, Elsa Marianna. 2024. *Piiblist pärit fraseologismide tundmine ja kasutus eesti keelt emakeelena ja teise keelena kõnelevate üliõpilaste hulgas*. Tartu Ülikool, eesti ja üldkeeleteaduse instituut. Bakalaureusetöö. https://hdl.handle.net/10062/102079.

Metsla, Triin. 2014. *Eesti keele fraseoloogia tõlkimisest Tõnu Õnnepalu „Piiririigi“ näitel*. Tartu Ülikool, eesti ja üldkeeleteaduse instituut. Bakalaureusetöö. http://hdl.handle.net/10062/43969.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: a corpus-based approach* (Oxford Studies in Lexicography and Lexicology). Oxford: Clarendon press.
Muischnek, Kadri & Heiki-Jaan Kaalep. 2009. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. *Eesti Rakenduslingvistika Ühingu aastaraamat* 5(0). 157–172. https://doi.org/10.5128/ERYa5.10.
Muischnek, Kadri & Liina Lindström. 2020. Digitaalsed tekstiandmed ja korpuslingvistika. *Kuidas mõista andmestunud maailma? Metodoloogiline teejuht* (Gigantum Humeris), 306–239. Tallinn: Tallinna ülikooli kirjastus.
Muischnek, Kadri, Kaili Müürisepp & Tiina Puolakainen. 2014. Dependency Parsing of Estonian: Statistical and Rule-based Approaches. *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/978-1-61499-442-8-111.
PARSEME. 2024. PARSEME corpora. *GitLab*. https://gitlab.com/parseme/corpora/-/wikis/home. (Vaadatud 21.02.2025).
PARSEME. s.a. PARSEME Shared Task 1.3 - Annotation guidelines (v. 2.0). https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=050_Tests_for_VERBAL_MWEs. (Vaadatud 09.05.2025).
Python Software Foundation. s.a. Python Language Reference, v.3. http://www.python.org. (Vaadatud 08.05.2025).
R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/. (Vaadatud 08.05.2025).
Ramisch, Carlos. 2023. *Multiword expressions in computational linguistics*. Aix Marseille Université. Habilitatsiooni väitekiri. https://theses.hal.science/tel-04216223v1.
Ramisch, Carlos & Aline Villavicencio. 2022. Computational Treatment of Multiword Expressions. Ruslan Mitkov (toim), *The Oxford handbook of computational linguistics*. 2a ed. Oxford New York: Oxford university press.
Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. Alexander Gelbukh (toim), *Computational Linguistics and Intelligent Text Processing* (Lecture Notes in Computer Science), kd 2276, 1–15. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45715-1_1.
Savary, Agata, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, jt. 2023. PARSEME corpus release 1.3. *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 24–35. Dubrovnik, Croatia: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.mwe-1.6.
Savary, Agata, Voula Giouli, Carlos Ramisch, Stella Markantonatou & Sara Stymne. *s. a.* (b). Nominal and modifier MWE tests and examples. *Google Docs*. https://docs.google.com/document/d/1dmjEFC8H-f6aiPVoTzJvpqUAy2SfOKbL7FAo2TU_tDI/edit?tab=t.0#heading=h.gt53hu7d9q5p. (Vaadatud 08.05.2025).
Savary, Agata, Voula Giouli, Carlos Ramisch, Stella Markantonatou & Sara Stymne. *s. a.* (a). Nominal and modifier MWE guidelines - a draft from Athens. *Google Docs*. https://docs.google.com/document/d/1bvjSwHpj8I2zJXmftCpx19u3BNWdKtdeg21f4YVHhWw/edit?usp=embed_facebook. (Vaadatud 13.07.2024).
Savary, Agata, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, jt. 2015. PARSEME -- PARSing and Multiword Expressions within a European multilingual network. *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*. Poola.

Savary, Agata, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch & Joakim Nivre. 2023. PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions. *Northern European Journal of Language Technology* 9(1). https://doi.org/10.3384/nejlt.2000-1533.2023.4453.

Svensson, Maria Helena. 2008. A very complex criterion of fixedness: Non-compositionality. Fanny Meunire & Sylviane Granger (toim), *Phraseology: an interdisciplinary perspective*, 81–93. John Benjamins Publishing Company. https://benjamins.com/catalog/z.139.

Sõnaveeb. 2025a. Pealinn. *EKI ühendsõnastik 2025*. Eesti Keele Instituut. https://sonaveeb.ee/search/unif/dlall/dsall/pealinn/1/est. (Vaadatud 18.05.2025).

Sõnaveeb. 2025b. Pea-. *EKI ühendsõnastik 2025*. Eesti Keele Instituut. https://sonaveeb.ee/search/unif/dlall/dsall/pea/6/est. (Vaadatud 18.05.2025).

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, jt. 2024. Universal Dependencies 2.14. http://hdl.handle.net/11234/1-5502. (Vaadatud 21.05.2025).

UD. *s. a.* Universal Dependencies. https://universaldependencies.org/. (Vaadatud 21.02.2025).

UniDive. *s. a.* Working Group 1: Corpus Annotation. *UniDive*. https://unidive.lisn.upsaclay.fr/doku.php?id=wg1:wg1. (Vaadatud 21.02.2025).

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis* (Use R!). 2nd ed. 2016. Springer-Verlag New York. https://doi.org/10.1007/978-3-319-24277-4.

Õim, Asta & Katre Õim. 2019. *Lähtekohti eesti fraseoloogia käsitlemiseks* (Eesti Teaduste Akadeemia Emakeele Seltsi toimetised nr. 76). Tallinn: EKSA.

Õunap, Reet. 2010. *Põhikooli- ja gümnaasiumiõpilaste fraseoloogiatundmine*. Tartu Ülikool, eesti ja üldkeeleteaduse instituut. Magistritöö. http://hdl.handle.net/10062/51640.

Nominal, Adjectival and Adverbial Multiword Expressions in Estonian Texts. Summary

The aim of this thesis was to find out which multiword expressions (MWEs) are the most frequent in Estonian texts, what is their sentence distribution (nominal, adjectival or adverbial), what form they take (written as one word or multiple words) and how they differ by text type (journalistic, fiction and popular science). An additional aim was to find out how tests (from the international UniDive initative) for identifying MWEs can be adapted for Estonian. The reasearch questions were the following:

1. Which multiword expressions and with what distribution appear the most in texts?
2. How does the use of multiword expressions differ by text type?
3. How can the multiword expression tests by the UniDive initiative be adapted for Estonian and which tests identify the most multiword expressions?
4. How many of the found multiword expressions are written as one word and how many of the compound words written as one word are multiword expressions?

To this end, I tagged nominal, adjective and adverbial MWEs in five hundred sentences for each text type from the Estonian Universal Dependencies tree bank (Zeman jt 2024). The MWEs were marked according to sentence distribution into noun phrases, whose subtypes were pronominal, nominal and deverbal noun phrases, and modifier phrases, whose subtypes were adjectival, adverbial and deverbal modifiers. For identifying MWEs, I used and adapted fifteen tests form the UniDive initiative for finding noun phrases and eight tests for finding modifier phrases. The tests were applied in a specific order until one of them received a positive result.

Among the nominal tests, the test for finding pronominal MWEs had to be adapted, because it was based on a previously existing list of pronominal MWEs, which does not exist in Estonian. Therefore, I also ran the remaining tests with pronominal MWE candidates but marked them with the pronominal test. In the future, it would be better to either make the required list in Estonian or treat nominal and pronominal MWEs similar to the modifier ones, where the tests are the same and the type is determined after receiving a positive test result. In the modifier tests, I added a few tests that were originally omitted, and therefore also changed the order a bit. In the future, it would be better to make the order of the additional tests more logical.

From the 18-thousand-word corpus, I found 691 multiword expressions, of which 472 were unique wordforms and 356 were unique lemmas. More than half of the MWEs were nominal,

less than a fifth were adverbials, and the rest types were less than a tenth of the expressions. 70% of the lemmas occurred only once and over 80% occurred in only one file out of nine, including two of the three most frequent MWEs.

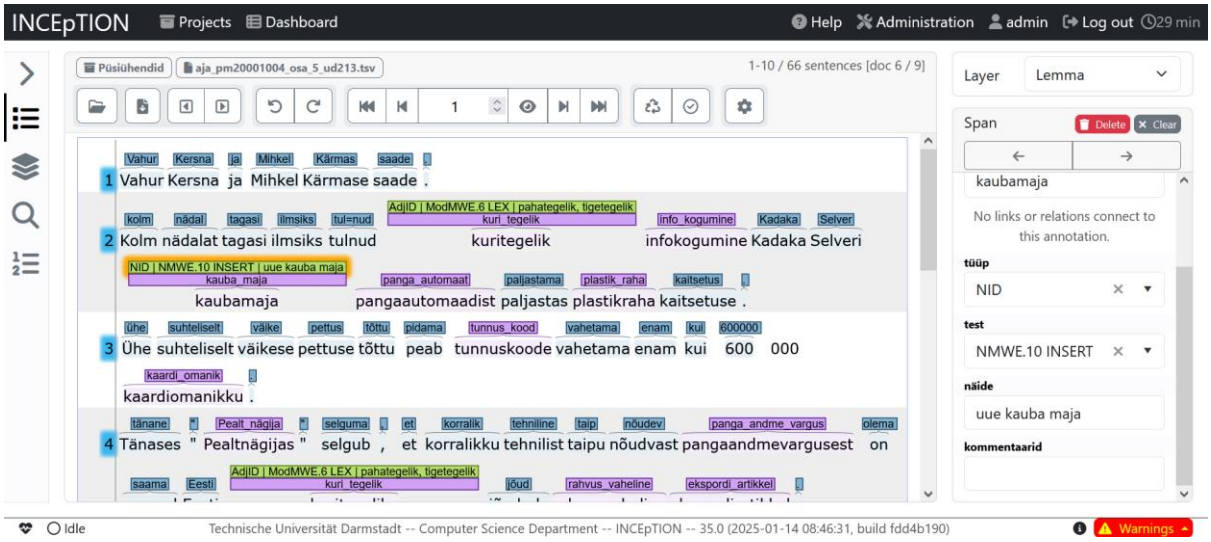
By text type, MWEs appeared more in popular science texts and less in fiction texts. However, journalistic texts had the biggest number of unique lemmas and there were the same amount of unique lemmas in popular science and fiction texts. Popular science texts may have had more repeated lemmas because the terminology was used more consistently there in comparison to journalistic texts, where terminology usage was more varied. Additionally, the proportion of modifier MWEs was higher in fiction texts.

Nominal MWEs were mostly identified by the modifier insertion test but also often by the lexical replacement test and the morphological change test, On the other hand, the modifier MWEs were identified most by the lexical replacement test and after that by the modifier insertion test. This is probably because it is difficult to find modifiers for modifiers. None of the nominal expressions found were identified by the syntactic modification test nor by the irregular internal syntactic structure tests, although there are examples of these in Estonian. None of the adjectival expressions were identified by the syntactic modification nor by the cranberry word test, for which it is also difficult to find examples in Estonian.

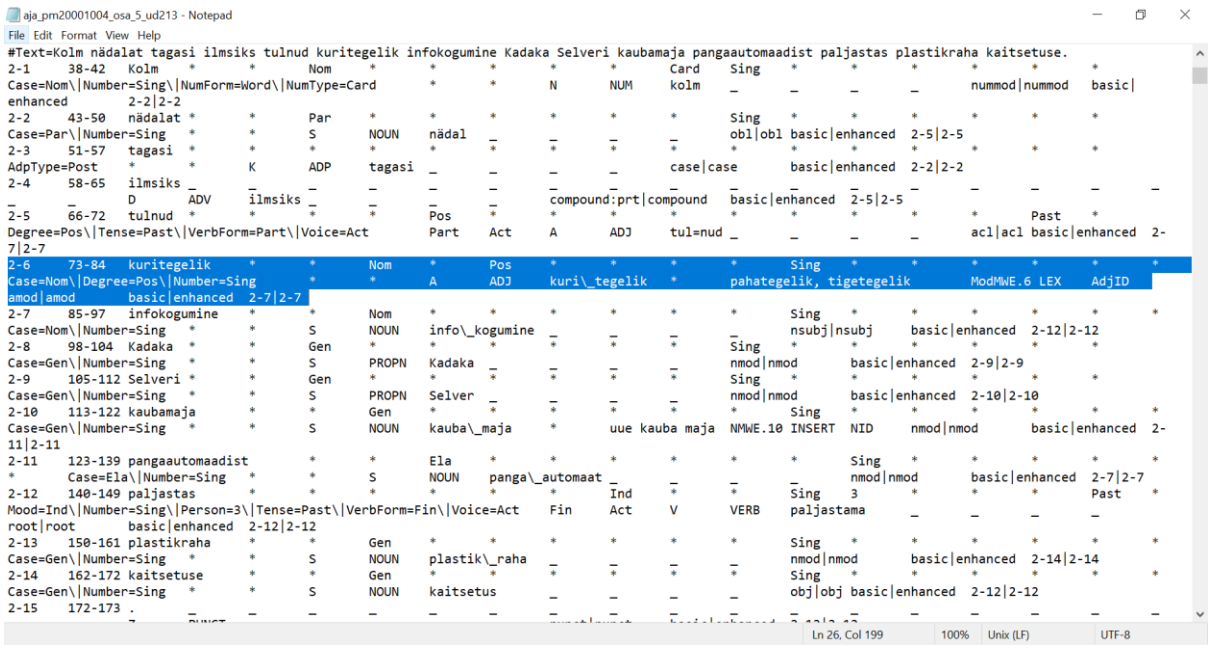
Almost 90% of the MWEs found were written as one word and 31% of the compound words that occurred were MWEs, although the proportion drops to 14% when looking at unique lemmas. There were more compound words in journalistic texts and all the pronominal MWEs and MWEs identified by three tests were written as one word. On the other hand, there were fewer compound words written as one word in the popular science tests in terms of general frequency. Additionally, all MWEs identified by the syntactic change test were written as separate words, though there were only two of those.

In the future, it would be interesting to study the same things using a larger amount of materjal and see if and how the results change. In addition, MWEs with other distributions, such as verbal and function phrases, could be included in a future study. It would also be interesting to compare the results with those from other languages based on the same framework.

Lisa 1. Ekraanipilt ühest märgendatud tekstist INCEpTIONis .



Lisa 2. Ekraanipilt ühest märgendatud tekstist tsv-faili kujul.



Lisa 3. Tabel vähemalt kaks korda esinenud püsiühenditega.

Lemmad	Vormid	Kokku	Aja	Ilu	Hor	Liik	Test	Näide
maailm	maailm, maailmalt, maailmas, maailmaga, maailma, maailmale, maailmast	40	16	5	19	NID	NMWE.10 INSERT	selle maa ilm
must auk	musta augu, mustad augud, musti auke, musta auguga, musta auku, must auk, mustade aukude, mustadest aukudest, mustast august	33	0	0	33	NID	NMWE.10 INSERT	üleni must auk
piirikoer	piirikoerast, piirikoera, piirikoer, piirikoerte, piirikoeri, piirikoerad, piirikoeral, piirikoerana	23	23	0	0	NID	NMWE.10 INSERT	selge/ühise piiri koer
peaaegu	peaaegu	18	4	9	5	AdvID	ModMWE.6 LEX	ligiaegu
vastuolu	vastuolud, vastuoludega, vastuolude, vastuoludest	10	0	0	10	NID	NMWE.10 INSERT	sellele vastu olu
olukord	olukord, olukorra, olukorras	9	5	1	3	NID	NMWE.10 INSERT	poliitilise olu kord
aastakümme	aastakümnetel, aastakümne, aastakümneid	8	1	0	7	NID	NMWE.8 IRREG_STRUCT	kümme aastat
eelkõige	eelkõige	8	2	0	6	AdvID	ModMWE.3 IRREG_STRUCT	kõige eel
kokkulepe	kokkuleppe, kokkulepe, kokkuleppega, kokkuleppele, kokkuleppel	8	0	3	5	VMWENom	NMWE.15 DEVERBAL	kokku leppima
unenägu	unenägudes, unenäod, unenäos, unenäost, unenäo, unenägu	8	0	8	0	NID	NMWE.10 INSERT	sügava une nägu
ajalugu	ajaloos, ajaloost, ajalugu, ajaloo	7	6	0	1	NID	NMWE.14 LEX	ajajutt
ülikool	ülikooli, ülikoolist, ülikool, ülikoolide	7	2	0	5	NID	NMWE.14 LEX	superkool
ülesanne	ülesanne, ülesannet, ülesandeid, ülesandeks	7	6	0	1	NID	NMWE.10 INSERT	kõige üles anne
esialgu	esialgu	6	0	0	6	AdvID	ModMWE.1 Cran	algu
maakera	maakera	6	0	0	6	NID	NMWE.7 MORPH	maakerad
eesmärk	eesmärk, eesmärgi, eesmärgiks	5	5	0	0	NID	NMWE.14 LEX	tagamärk
maantee	maanteel, maanteele, maanteed, maanteeni	5	1	4	0	NID	NMWE.6 CRAN	maan
suurriik	suurriigiga, suurriike, suurriikide, suurriigid	5	0	0	5	NID	NMWE.10 INSERT	väga suur riik
otsekui	otsekui	5	3	2	0	AdvID	ModMWE.5 INSERT	ainult/väga otse kui
laulutaat	laulutaati, laulutaat, laulutaadi, laulutaadile	5	0	5	0	NID	NMWE.10 INSERT	keerulise laulu taat
igaüks	igaüks, igaühes	4	1	1	2	PronID	NMWE.1 IS_PRON	*

tänapäev	tänapäeval, tänapäeva, tänapäevaks, tänapäevalgi	4	2	0	2	NID	NMWE.7 MORPH	tänapäevad
ajaleht	ajaleht, ajalehte, ajalehes, ajalehe	4	2	2	0	NID	NMWE.10 INSERT	pika aja leht
vastasseis	vastasseis, vastasseisust	4	1	0	3	NID	NMWE.10 INSERT	maja vastas seis
raudteelane	raudteelaste, raudteelasi, raudteelased, raudteelastele	4	4	0	0	NID	NMWE.14 LEX	kuldteelane
vabariik	vabariigi	4	2	0	2	NID	NMWE.10 INSERT	täiesti vaba riik
välismaa	välismaa, välismaalt, välismaal	4	3	0	1	NID	NMWE.14 LEX	sisemaa
tagajärg	tagajärgedest, tagajärgedele, tagajärjed, tagajärgi	4	0	0	4	NID	NMWE.10 INSERT	kõige taga järg
kasvuhoone	kasvuhoone, kasvuhoonega	4	0	0	4	NID	NMWE.10 INSERT	kiire kasvu hoone
peremees	peremehe, peremees, peremehega	4	3	1	0	NID	NMWE.10 INSERT	selle pere mees
tugitool	tugitool, tugitooli	4	0	4	0	NID	NMWE.14 LEX	toetustool, tugiiste
helilooja	helilooja, heliloojate	3	3	0	0	NID	NMWE.10 INSERT	kõrge heli looja
tõenäoliselt	tõenäoliselt	3	2	1	0	AdvID	ModMWE.6 LEX	tõeilmeliselt, tõekäeliselt
kuritegu	kuritegu, kuritegude, kuriteo	3	3	0	0	NID	NMWE.14 LEX	pahategu, tigetegu
ülemineku	ülemineku	3	3	0	0	VMWENom	NMWE.15 DEVERBAL	üle minema
ajakiri	ajakirjas, ajakiri	3	1	0	2	NID	NMWE.10 INSERT	pika aja kiri
kivisüsi	kivisöe, kivisöeks	3	2	0	1	NID	NMWE.7 MORPH	kivisöed
tähelepanu	tähelepanu	3	1	0	2	VMWENom	NMWE.15 DEVERBAL	tähele panema
ajalooline	ajaloolist, ajaloolised	3	1	0	2	AdjID	ModMWE.6 LEX	ajajutuline
nimekiri	nimekirja	3	2	1	0	NID	NMWE.10 INSERT	uue nime kiri
ülikool	ülikooli, ülikoolis	3	0	0	3	NID	NMWE.14 LEX	superkool
tõepoolest	tõepoolest	3	1	1	1	AdvID	ModMWE.5 INSERT	algse tõe poolest
seisukoht	seisukohalt, seisukohal	3	0	0	3	NID	NMWE.10 INSERT	sirgelt seisu kohalt
poolsaar	poolsaare, poolsaarel	3	1	0	2	NID	NMWE.10 INSERT	ainult pool saar
seepärast	seepärast	3	1	1	1	AdvID	ModMWE.3 IRREG_STRUCT	selle pärast

omaette	omaette	3	0	2	1	AdvID	ModMWE.6 LEX	omataha
mitmekesisus	mitmekesisuse	3	0	0	3	NID	NMWE.7 MORPH	mitmekesisused
uudishimu	uudishimu	3	0	2	1	NID	NMWE.7 MORPH	uudishimud
kasvuhoonegaas	kasvuhoonegaasid, kasvuhoonegaaside, kasvuhoonegaase	3	0	0	3	NID	NMWE.10 INSERT	väikese kasvuhoone gaasid
piiririkkuja	piiririkkuja, piiririkkujat, piiririkkujad	3	3	0	0	VMWENom	NMWE.15 DEVERBAL	piiri rikkuma
elukutse	elukutse, elukutsega	3	3	0	0	NID	NMWE.10 INSERT	tema/pika elu kutse
võibolla	võibolla	3	1	2	0	AdvID	ModMWE.8 IRREG_STRUCT_DISTRIB	verbifraas
kahtlusalune	kahtlusaluse, kahtlusalusest, kahtlusalune	3	3	0	0	NID	NMWE.14 LEX	kahtluspealne
nagunii	nagunii	3	0	3	0	AdvID	ModMWE.6 LEX	nagunõnda
vaatepilt	vaatepilti, vaatepilt	3	0	3	0	NID	NMWE.12 COORD	ekraani- ja vaatepilt
abikaasa	abikaasa	2	2	0	0	NID	NMWE.10 INSERT	suure/rahalise abi kaasa
niivõrd	niivõrd	2	1	1	0	AdvID	ModMWE.6 LEX	nõndavõrd
suurepärane	suurepärane	2	2	0	0	AdjID	ModMWE.6 LEX	väikesepärane
keskaeg	keskajal	2	1	0	1	NID	NMWE.7 MORPH	keskajad
läänemaailm	läänemaailmale, läänemaailmas	2	1	0	1	NID	NMWE.7 MORPH	läänemaailmad
kuupäev	kuupäeva, kuupäeval	2	1	0	1	NID	NMWE.10 INSERT	eelmise kuu päev
maavärin	maavärin, maavärinat	2	2	0	0	NID	NMWE.14 LEX	maavõbin
tuletõrjuja	tuletõrjujad, tuletõrjujalt	2	2	0	0	NID	NMWE.10 INSERT	suure tule tõrjuja
koosseis	koosseisus, koosseisu	2	1	1	0	NID	NMWE.10 INSERT	teistega koos seis
kodusõda	kodusõda	2	1	0	1	NID	NMWE.10 INSERT	minu kodu sõda
ainuüksi	ainuüksi	2	1	0	1	AdvID	ModMWE.3 IRREG_STRUCT	vb ainult üksi
elumaja	elumaja	2	2	0	0	NID	NMWE.10 INSERT	senise/avaliku elu maja

meesterahvas	meesterahvast, meesterahva	2	2	0	0	NID	NMWE.10 INSERT	nende meeste rahvas
hooaeg	hooaja, hooaega	2	2	0	0	NID	NMWE.10 INSERT	väikese hoo aeg
pealtvaataja	pealtvaataja	2	1	1	0	VM WENom	NMWE.15 DEVERBAL	pealt vaatama
auhind	auhinna, auhinnale	2	2	0	0	NID	NMWE.10 INSERT	suure au hind
päevakord	päevakorda	2	0	0	2	NID	NMWE.10 INSERT	homse päeva kord
ülevaade	ülevaate, ülevaadet	2	0	0	2	VM WENom	NMWE.15 DEVERBAL	üle vaatama
läbimõõt	läbimõõduga, läbimõõt	2	0	0	2	NID	NMWE.10 INSERT	rõngast läbi mõõt
viimnepäev	viimsepäeva, viimnepäev	2	0	1	1	NID	NMWE.10 INSERT	suve viimne päev
üheaegselt	üheaegselt	2	0	0	2	AdvID	ModMWE.6 LEX	kaheaegselt
läbirääkimine	läbirääkimistel, läbirääkimised	2	0	0	2	VM WENom	NMWE.15 DEVERBAL	läbi rääkima
kallaletung	kallaletung	2	0	0	2	VM WENom	NMWE.15 DEVERBAL	kallale tungima
eeskätt	eeskätt	2	0	0	2	AdvID	ModMWE.2 MORPH	eeskäsi
ülestõus	ülestõus, ülestõusu	2	0	0	2	VM WENom	NMWE.15 DEVERBAL	üles tõusma
ettepanek	ettepaneku, ettepanekuid	2	0	0	2	VM WENom	NMWE.15 DEVERBAL	ette panema
julgeolek	julgeoleku	2	1	0	1	NID	NMWE.10 INSERT	täiesti julge olek
rannajoon	rannajoone, rannajoon	2	0	1	1	NID	NMWE.10 INSERT	liivase ranna joon
tulipunkt	tulipunkti	2	0	0	2	NID	NMWE.14 LEX	leegipunkt
parasjagu	parasjagu	2	0	1	1	AdvID	ModMWE.5 INSERT	täpselt paras jagu
ligilähedaselt	ligilähedaseltki	2	0	0	2	AdvID	ModMWE.6 LEX	pealähedaselt
pruunkaru	pruunkaru	2	0	0	2	NID	NMWE.14 LEX	mustkaru, hallkaru
ükskord	ükskord	2	0	1	1	AdvID	ModMWE.5 INSERT	ainult üks kord

ükskõik	ükskõik	2	1	1	0	AdvID	ModMWE.5 INSERT	ainult üks kõik
kurjategija	kurjategijad, kurjategija	2	2	0	0	NID	NMWE.14 LEX	halvategija, kurjusetegija
kogu aeg	kogu aeg	2	1	1	0	AdvID	ModMWE.6 LEX	terve aeg
metsajälg	metsajäljes, metsajälg	2	2	0	0	NID	NMWE.7 MORPH	metsajäljed
asjaajamine	asjaajamine, asjaajamisest	2	1	1	0	VMWENom	NMWE.15 DEVERBAL	asja ajama
süvamuusika	süvamuusika	2	2	0	0	NID	NMWE.14 LEX	pinnamusika
kuritegelik	kuritegelik, kuritegelike	2	2	0	0	AdjID	ModMWE.6 LEX	pahategelik, tigetegelik
palverändur	palverändur	2	0	2	0	NID	NMWE.10 INSERT	alandlikupalverändur
keskaegne	keskaegne, keskaegses	2	0	2	0	AdjID	ModMWE.5 INSERT	täpselt keskaegne
rahvakunstnik	rahvakunstnikele, rahvakunstniku	2	0	2	0	NID	NMWE.10 INSERT	selle rahva kunstnik
ülekohtune	ülekohtune	2	0	2	0	AdjID	ModMWE.6 LEX	alakohtune
kõrvulukustav	kõrvulukustava, kõrvulukustav	2	0	2	0	AdjID	ModMWE.6 LEX	silmilukustav
iseenesest	iseenesest	2	0	2	0	AdvID	ModMWE.5 INSERT	täiesti iseenesest
üksisilmi	üksisilmi	2	0	2	0	AdvID	ModMWE.6 LEX	kaksisilmi, üksikõrvu
kõrvulukustavalt	kõrvulukustavalt	2	0	2	0	AdvID	ModMWE.6 LEX	silmilukustavalt
käsivars	käsivars, käsivarte	2	0	2	0	NID	NMWE.12 COORD	käsi- ja roosivars
koosolek	koosolekute, koosolekul	2	0	2	0	NID	NMWE.10 INSERT	teisega koos olek
kättemaks	kättemaksu, kättemaksuks	2	0	2	0	NID	NMWE.12 COORD	kätte- ja tollimaks

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Karoliina Jõgi,

1) annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Nimi-, omadus- ja määrsõnafraasilised püsiühendid eestikeelsetes tekstides“, mille juhendaja on Kadri Muischnek, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2) annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi 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.

Allkirjastatud digitaalselt

24.05.2025