Emakeele Seltsi aastaraamat 69 (2023), 9–34
http://dx.doi.org/10.3176/esa69.01

KEELEKASUTUSREEGLITE TULETAMINE JA VEATUVASTUS MÄÄRSÕNA SISALDAVATE SÕNALIIGIJÄRJENDITE NÄITEL

KAIS ALLKIVI-METSOJA,
PILLE ESLON, JAAGUP KIPPAR

Annotatsioon. Tutvustame uut tarkvara, mis võimaldab sõnaliigijärjendite esinemiskonteksti alusel tuletada eesti keele kasutusreegleid ja tuvastada grammatikavigu. Nii on ka artikli eesmärk kahetine. Tuginedes eesti keele koondkorpuse põhjal loodud statistilisele keelemudelile ja keskendudes määrsõna sisaldavatele kolmesõnalistele järjenditele, 1) anname ülevaate sõnaliigikooslustest, mida eelistatakse kasutada lause alguses ja lõpus; 2) kirjeldame veakohti, mis tulevad sõnajärjendite ebatõenäolise kasutuskonteksti järgi esile eesti keele kui teise keele õppijate tekstiloomes. Leitud vead on sagedamini seotud V2-sõnajärjega (osa)lause alguses ja määruse paigutusega lause lõpus, valitud meetod aitab avastada ka puuduvaid ja liigseid sõnu. Normipärase ja ebatüüpilise keelekasutuse kombineeritud analüüs annab ainest nii automaatse veatuvastuse täpsustamise kui ka veaparanduste soovitamise jaoks.

Võtmesõnad: grammatikavigade tuvastus, keeletöötlus, morfosüntaks, kasutuspõhine keelekirjeldus, n-grammid

1. Sissejuhatus

Uurimus seostub eesti keele automaatkorrektuuri vahendite arendusega[1]. Veaparanduse tarbeks on kättesaadavaks tehtud neuromasintõlkel põhinevad grammatikakontrolli ja statistilised õigekirjakontrolli mudelid (Allkivi-Metsoja, Kippar 2023; Luhtaru jt 2024). Grammatikakontrollijaga on saadud häid tulemusi täheortograafia, puuduvate kirjavahemärkide,

[1] Tööd on toetatud HTMi „Eesti keeletehnoloogia“ programmist (EKTB25 „Eestikeelse teksti automaatkorrektuur“).

käänd- ja tegusõna vormivigade, kokku- ja lahkukirjutuse parandamisel, kuid sõnajärje- ja sõnavalikuvigadest ning liigsetest-puuduvatest sõnadest parandab praegune parim korrektuurimudel umbes poole[2]. Nende vealiikide tuvastamisele saab kaasa aidata rakendus, mis toob tekstis esile vähetõenäolised sõnaliigijärjendid.

Tekstis regulaarselt esinevad sõnaliigi n-grammid (sõnaliigijärjendid) moodustavad keelekasutusmustreid, mis eristuvad sõnaühenditest selle poolest, et neil on a) püsikindel struktuur, nt kolm järjestikust määrsõna (nt *kahjuks ka palju*); b) stereotüüpne/tüüpiline vormistus ja tekstifunktsioonid, nt kahe järjestikuse määrsõna ja sidesõna puhul kasutatakse vastandavat paarissidendit *ei .. ega* (*ei rohkem ega vähem*); c) piiritletud leksikaalgrammatiline varieeruvus, nt tegusõna kasutamisel koos kahe järgneva määrsõnaga moodustuvad hoogsat tegevust tähistavad ütlused (*sammume/liigume/põrutame kindlalt edasi*) ja piltlikud väljendid (*ei vea enam välja*).

Sõnaliikide n-gramme (ingl *POS-grams*) on lingvistikas analüüsitud mitmetel eesmärkidel, mh fraseologismide tuvastamiseks (Brett, Pinna 2015), tekstide klassifitseerimiseks (Kapusta jt 2021), korduvate grammatiliste struktuuride leidmiseks keeleõppe vajadustel (Cappelle, Grabar 2016) ja veatuvastuses. Jian-cheng Wu jt (2013) on n-grammide alusel määranud ingliskeelsetes lausetes tegusõnale eelneva kaassõna valiku vigu. Jahangir Md. Alam jt (2007) on kasutanud n-grammide tõenäosusi bangla- ja ingliskeelsete lausete grammatilise korrektsuse hindamiseks. Mikko Aulamo (2019) on soome keele grammatikavigu tuvastanud 2–5 komponendist koosnevate sõnaliigi- ja grammatiliste vormide järjendite sageduse põhjal. Inglise keeles on valdavalt analüüsitud sõnajärjendeid, sagedamini vaadeldakse 3–4-sõnalisi üksusi (vt De Cock, Granger 2021).

Eesti keelele loodud tarkvara nimetame sõnaliigijärjendite leidjaks[3]. See koosneb kahest komponendist.

1. Esimene programm[4] eraldab tekstimaterjalist trigrammid (sõnaliigikolmikud), arvutab nende kasutuskontekstide sageduse ja osakaalu trigrammi ees (lause algus või eelnev sõnaliik) ja järel

[2] https://github.com/TartuNLP/grammar-worker/tree/main/models.

[3] Huvilised saavad sõnaliigijärjendite leidmist ja veatuvastust katsetada Google Colaboratory demolehtedel (Sõnaliigijärjendite leidja 2024; Sõnaliigijärjenditel põhinev veatuvastus 2024).

[4] https://github.com/tlu-dt-nlp/POSgram-contexts.

(lause lõpp või järgnev sõnaliik). Analüüsides selliselt mahukat ja võrdlemisi esinduslikku üldkeelekorpust, saab koostada statistilise keelemudeli, mis võimaldab konteksti põhjal tuvastada tüüpilisi ja ebatüüpilisi sõnaliigijärjendeid.

2. Teine programm[5] otsib keelemudelile tuginedes tekstist vähetõenäolisi sõnaliigijärjendeid, mille ees- või järelkonteksti tõenäosus jääb keelemudeli alusel alla kasutaja seatud piiri (nt 5%). Nii toob otsing lause kaupa esile trigrammid, mida tarvitatakse harva lause alguses/lõpus või mis pole ootuspärased eelneva/järgneva sõnaliigi tõttu.

Et mõlema programmi kasutust demonstreerida, analüüsime artiklis kahesugust keeleainest: 1) eesti keele koondkorpuse põhjal eritleme trigrammide ümber eelistatud kasutuskontekste; 2) võrdluses koondkorpusega tuvastame eesti keelt teise keelena õppijate tekstidest ebaharilikus kontekstis esinevaid trigramme ja kõrvutame neid veamärgendusega. Oleme materjali piiranud määrsõna sisaldavate trigrammidega. Täpsemalt keskendume järjenditele, mida eelistatakse kasutada lause algul või lõpus. Veanäidete puhul vaatleme ka lausesisest konteksti.

Uurimus on korpusest tulenev ja jätkab lingvistilise klasteranalüüsi suunda (vt Eslon, Allkivi-Metsoja 2018), kuid sõnaliigijärjendite leidja erineb mitmeti varem loodud Klastrileidjast[6] ja selle arendusest Mustrileidjast[7]: hetkel otsib uus rakendus üksnes sõnaliigi n-gramme, mitte vormi- ja lauseliikmete järjendeid; grammatiline märgendus ei põhine EstCG reeglipõhisel süntaksianalüsaatoril; uudsed on andmed sõnaliikide varieerumisest n-grammi ees/järel ja veatuvastus nende esinemisstatistika alusel.

Oleme seisukohal, et sõnaliigi n-grammi struktuuri on kodeeritud selle komponentide kooskasutuse fonoloogilised, süntaktilised ja kontseptuaalsed reeglid, mis aktiveeruvad keelekasutuses semantika-grammatika piirimail (vt Jackendoff 2017).

---

[5] https://github.com/tlu-dt-nlp/POSgram-errors.

[6] https://evkk.tlu.ee/vers1/Search/search_reeglid.html.

[7] https://elle.tlu.ee/tools/clusterfinder.

2. Keelematerjal ja sõnaliigijärjendite tuvastamine

Sõnaliigijärjendite leidja rakendab sõna- ja lausepiiride määramiseks ning sõnaliigituvastuseks keeletöötluspaketti Stanza (Qi jt 2020). Stanfordi ülikoolis loodud tarkvara põhineb närvivõrgumudelitel, mida on eestikeelsete tekstide analüüsimiseks treenitud Eesti universaalsete sõltuvustega märgendatud puudepanga (EstUD) versiooniga 2.12[8]. Stanza on saavutanud EstBERTi ja RoBERTa keelemudelitest paremaid tulemusi sõnestamises ja sõnaliikide märgendamises, kuid jäänud lausestamises alla RoBERTale, nõudes samas vähem arvutusressurssi (EstSpacy 2021). Teisalt on Stanza EstUD tuvastanud kirjakeelsete tekstide lausepiire täpsemini kui PipeUD mudel ja EstNLTK reeglipõhine märgendaja Vabamorf, ehkki viimane on andnud paremaid tulemusi veebitekstidega (Sirts, Peekman 2020).

Sisendteksti töötlus toimub lause kaupa. Iga lause sõnaliigid kirjutatakse vastuste faili, millest on võimalik teha päringuid. Sõnaliigituvastuses oleme lähtunud CoNLL-U märgendusformaadi[9] atribuudist *xpos* ehk keelespetsiifilistest sõnaliigimärgenditest[10]. Lisaks võtsime kasutusele lause alguse (^) ja lõpu ($) sümboli, mis tähistavad trigrammi ees- või järelkonteksti puudumist (vt tabel 1). Edasiseks analüüsiks viib tööriist laused sõnaliigijärjendi kujule, nt lause *Juku ja Kati laulsid valjusti* esitus on ^SJSVD$ (nimisõna-sidesõna-nimisõna-tegusõna-määrsõna). Praegune versioon ei arvesta järjendite osana kirjavahemärke (märgend *Z*), sest nii saab paremini vaadelda sõnade järgnevusi lauses, mh osalausepiiril. Siin ja edaspidi kasutame tabelis 1 antud sõnaliigitähiseid, noolsulg > eraldab trigrammist eeskonteksti ja < järelkonteksti. Kõnealune sõnakolmik (lause alguses/lõpus) või -nelik (lause sees) on keelenäidetes alla joonitud.

Trigrammide kasutuskontekstide leidmiseks eraldatakse lausetest nelja märgendi järjendid (tetragrammid). Eelnevas näitelauses on neid neli: ^SJS – *Juku ja Kati*, SJSV – *Juku ja Kati laulsid*, JSVD – *ja Kati laulsid valjusti*, SVD$ – *Kati laulsid valjusti*. Et teha kindlaks trigrammide eeskontekstid, rühmitatakse tetragrammid kolme viimase märgendi järgi. Seejärel saab arvutada iga eristunud kolmiku eeskontekstide sageduse ja osakaalu kõigi selle kolmiku keelenäidete suhtes ning kolmiku enese

[8] https://stanfordnlp.github.io/stanza/available_models.html.

[9] https://universaldependencies.org/format.html.

[10] https://www.cl.ut.ee/korpused/morfkorpus/seletus/index.php?lang=et.

Tabel 1. Sõnaliigijärjendite tuvastamiseks kasutatavad märgendid

Märgend	Seletus
^	lause algus
A	omadussõna
D	määrsõna
G	omastavaline täiend
I	hüüdsõna
J	sidesõna
K	kaassõna
N	arvsõna
P	asesõna
S	nimisõna
V	tegusõna
X	abimäärsõna
Y	lühend
$	lause lõpp

sageduse ja osakaalu teiste kolmikute suhtes. Analoogselt, kuid vastupidises suunas toimib järelkontekstide otsing.

Kirjeldatud viisil koostasime statistilise keelemudeli, mis kajastab trigrammide ees- ja järelkontekstide tõenäosusi. Allikmaterjaliks võtsime eesti keele koondkorpuse[11] versiooni, mis on kättesaadav eesti keele ühendkorpuse alamkorpusena (u 181 mln sõna ja 13,2 mln lauset). Koondkorpus sisaldab peamiselt aja-, ilukirjandus-, teadus- ja seadustekste. Eeldasime, et see esindab veebikorpustega võrreldes standardsemat keelekasutust, mis võiks paremini sobida kirjalike tekstide veatuvastuse aluseks.

Leidsime koondkorpuse tekstidest 152,6 miljonit trigrammi. Siinse uurimuse jaoks sõelusime neist välja määrsõna sisaldavad, mida oli 49,3 miljonit (osakaal 32,3%). Neid sõnaliigijärgnevuse alusel ühendades jäi alles 419 trigrammi, mille hulgast eemaldasime tõlgendamatut komponenti (märgend *T*) sisaldavad ja need, mille osakaal korpuses on alla 0,0001%. Nii kitsenes järjendite arv 303-ni (osakaal korpuses kuni 1,610%). Artikli 3. peatükis piirdume näitega määrsõna sisaldavatest sõnaliigijärjenditest

[11] https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et.

lause alguses ja lõpus, kirjeldame sõnaliikide varieeruvust nende ees/järel ning anname lühiülevaate analüüsitud mustrite keelekasutusest. Me ei ole kõrvale jätnud ega parandanud juhtumeid, kus sõnaliigituvastus loeb määrsõnaks (*D*) abimäärsõna, kuna selliste olukordadega peab arvestama ka sõnaliigijärjendite leidja töös.

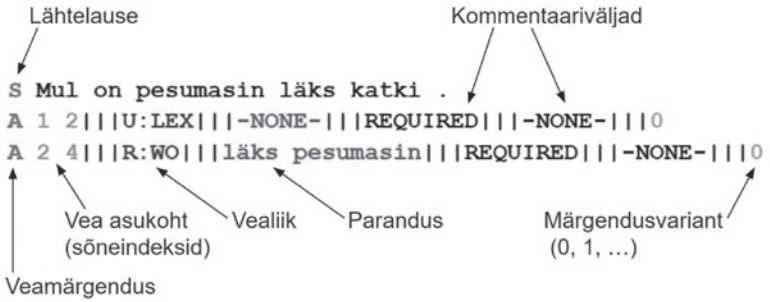
Veatuvastus keskendub ebatüüpilises kontekstis esinevatele sõnaliigijärjenditele. Programm tugineb koondkorpuse alusel koostatud keelemudelile. Kasutaja saab seada tõenäosuspiiri, millest väiksema osakaaluga kontekstid veaohtlikuks märgitakse. Vaikimisi oleme selleks määranud 5%. Kui protsendipiiri suurendada, suureneb ka veatuvastuse saagis ehk leitud vigade osakaal. Saagis võib aga lõputult kasvada, kui üha rohkem sõnajärjendeid on tähistatud võimaliku veana. Oleme kindlaks teinud, et 5% juures ei vähene tulemuste täpsus (õigustatud veatuvastuste osakaal) veel olulisel määral, nt võrreldes 2% piiriga.

Veatuvastuse täpsust ja saagist võimaldab hinnata eesti keele automaatkorrektuuri vahendite testimiseks loodud veamärgendusega korpus, mis koosneb eesti keele kui teise keele õppijate tekstidest (EstGEC-L2 2023). Testmaterjal on pärit eesti vahekeele korpusest[12] ja hõlmab 2029 lauset, jaotudes suhteliselt võrdselt nelja keeleoskustaseme (A2, B1, B2, C1) vahel. Vigade asukoha, liigi ja paranduse tähistamiseks on kasutusel M2-märgendusformaat[13]. Vealiigitus on eesti keele jaoks kohandatud ERRANTi klassifikatsiooni alusel (Bryant jt 2017). Üldisem liigitus tugineb sellele, kas sõna või kirjavahemärk on liigne, puudub või vajab asendamist. Neile vealiikidele vastab märgendi algul *U* (ingl *unnecessary*), *M* (*missing*) või *R* (*replacement*). Sõnaasenduste hulka kuuluvad täheortograafia-, algustähe-, kokku-lahkukirjutuse, vormi- ja sõnavaliku ning sõnajärjevead (asendatakse mitmesõnaline üksus). Kui ebaloomuliku sõnajärjega lauseosa hõlmab ka sõnatasandi vigu, on need märgitud omaette eksimustena, et nende üle saaks eraldi arvestust pidada.

Joonisel 1 on kujutatud veamärgenduse näide. Lausele on iga vea kohta lisatud märgendusrida. Selle alguses on sõneindeksid, mille vahel vigane üksus paikneb. Esimese sõne asukoht on 0–1, teisel sõnel 1–2 jne. Siinses näites on teine sõne märgendatud liigsena, kolmas ja neljas aga sõnajärjevea tõttu ära vahetatud: *Mul läks pesumasin katki*. Ühel lausel on kuni kolm märgendusversiooni (nummerdatud 0, 1 või 2 rea lõpus).

[12] https://elle.tlu.ee/tools.

[13] https://github.com/nusnlp/m2scorer.



Joonis 1. Veamärgenduse näide M2-formaadis

Kuigi veatuvastajaga saab otsida mis tahes vähetõenäolisi sõnaliigijärjendeid, käsitleme 4. peatükis üksnes määrsõna sisaldavate trigrammidega esile tulnud ebatäpset keelekasutust ja tarbetuid veatuvastusi.

3. Sõnaliigijärjendid lause alguses ja lõpus

Sõnaliigijärjendite esinemiskontekstide analüüs võiks keelekasutuse seaduspärade kirjelduse kõrval pakkuda sisendit automaatse veaparanduse jaoks, aidates kasutajale selgitada ebatavalistes sõnaliigijärjendites avalduvaid keelevigu ja pakkuda paremaid sõnastusi.

3.1. Lausealgulised sõnaliigijärjendid

Statistiline ülevaade. Määrsõna sisaldavaid trigramme, mille kasutuskontekstidest on sagedaim lause algus, leidus koondkorpuses 26. Nende osakaal jääb vahemikku 0,0001–0,845% (vt tabel 2). Kõige laiema levikuga järjend DVS esineb lause algul üle kolmandiku juhtudest. Tunduvalt harvem varieeruvad selle ees nimi- (19,21%), tegu- (15,35%), määr- (11,19%) ja sidesõna (10,43%), nt *Ministeerium võib-olla aitab juhtumit lahendada* (S>DVS); *Öelge, millal toimub kohtumine linnavalitsuse inimestega* (V>DVS); *Juba üleeile õnnitles peatreenerit .. Eesti Olümpiakomitee* (D>DVS); *Ja muuhulgas otsustati koosolekul taastada ..* (J>DVS).

Alati ei ole trigrammi lausealguline kasutus tugevalt eelistatud. Nii on DVD esinemus selles positsioonis suurim, kuid vahe ülejäänud eeskontekstidega pole kuigi oluline. On tõenäoline, et mustrit kasutatakse ka lause sees, kus talle võib eelneda nimisõna (23,02%; *Turg samuti määrab*

Tabel 2. Lausealgulised sõnaliigijärjendid reastatuna lause algul esinemise tõenäosuse järgi

Jrk	Järjend	Osakaal koond-korpuses %	Osakaal lause alguses %	Näide
1.	IJD	0,0007	72,11	*Ah et ammu* kestab juba?
2.	IAD	0,0001	69,87	*No hea küll*, müüvad ameeriklastele.
3.	IDV	0,0023	67,12	*Vaat nii arvan* mina!
4.	IDJ	0,0002	66,15	*No täpselt nagu* deodorandipudel!
5.	IID	0,0002	66,09	*Oh issand küll*.
6.	IDP	0,002	65,48	*Vaat seal sa* tahad istuda!
7.	IVD	0,0015	64,78	*No aidake ometigi*, ma suren!
8.	IDS	0,0008	63,64	*No näiteks ämbreid* pole mõtet .. müüagi.
9.	IPD	0,0007	61,28	*Säh sulle siis* praimingut.
10.	IDD	0,0012	60,87	*Vat nii siis*.
11.	DVX	0,0001	57,58	*Seejuures unustati tähele* panemata, et ..
12.	IDN	0,0001	55,78	*No vähemalt kolme* hulka pidime ikka pääsema.
13.	IDA	0,0007	50,7	*Oh kui raske* on ikkagi olla Isa!
14.	ISD	0,0005	44,96	*Oi jumal küll*!
15.	DVY	0,021	37,5	*Viimati lõppes MM* Argentinale .. õnnetult ..
16.	DVS	0,845	35,43	*Samuti tuleb kinnistu* hoida .. heas seisukorras.
17.	DVP	0,386	34,08	*Esiteks on see* lähemal.
18.	JID	0,0006	34,0	*Ja vat siis* sööme vastlakukleid kah!
19.	JDI	0,0001	33,51	*Ja nüüd, kurat*, see teepikendus!
20.	DVG	0,0043	29,76	*Paremini oskavad eesti* keelt Eesti .. muulased.
21.	DVA	0,204	27,34	*Öösel puhub nõrk* muutlik tuul ..
22.	PVD	0,432	27,23	*Tal on ju* lähedased sugulased Narvas.
23.	DVD	0,363	26,22	*Tänavu jätkatakse ka* .. puiestee asfalteerimist.
24.	DPV	0,185	24,17	*Eks seegi ole* fakt.
25.	DKV	0,0011	24,02	*Sealt alates on* usaldus .. jõuliselt suurenenud.
26.	AVD	0,066	21,0	*Positiivsena saab veel* välja tuua ..

nii kauba hinna kui ka käibe), harvem tegusõna (18,32%; *Nad hakkavad tegelikult ehitama ikka uut maja ..*). Ka DPV ees võib vabalt olla nimi- (23,15%; *Kodus ju seda pole näha*) või tegusõna (19,8%; *Lubage ka mul osaleda*), PVD ees nimisõna (26,74%; *Kontoris ma pole kunagi kohvita jäänud*). AVD kasutus lause algul hõlmab vaid viiendiku keelenäidetest ning sama edukalt võivad talle eelneda nii tegu- (20,85%; *Hinnata positiivsena saab ka esinemiskindlust*) kui ka määrsõna (20,7%; *Isegi peamine on mitte langeda liialdustesse*). Niisiis on nimetatud sõnaliigijärjendite esinemus lause alguses vaid kõige tõenäolisem võimalus. Samas, 13 trigrammi esinemus selles positsioonis jääb 50,7–72,11% vahele, kuid korpuses tervikuna on need väga harvad (osakaal 0,0001–0,0023%).

Lausealgulistele trigrammidele järgneb kas nimi- või tegusõna (vt tabel 3). Regulaarselt (üle 50% juhtudest) kasutatakse nimisõna nelja ja tegusõna ühe trigrammi järel: DVG<S, nt *Üldse peab vene köögi puhul .. arvestama*; DVA<S, nt *Samuti sätestab uus seadus, et ..* ; IDN<S, nt *No vähemalt kolme hulka pidime .. pääsema*; DVY<S, nt *Nüüd hakkab USA Kongress uurima ..* ; DVX<V, nt *Kas oled tähele pannud, et ..* .

Veidi väiksem esinemus (kolmandik või enam juhtudest) on nimisõnal kuue ja tegusõnal seitsme trigrammi järel. Niisuguseid järjendeid on 13.

a) Nimisõnaga: DVP<S, nt *Väliselt iseloomustab selliseid raamatuid, et ..* ; DVS<S, nt *Piirkonniti oli töötuse määr 2001. aasta IV ..* ; DKV<S, nt

Tabel 3. Lausealguliste sõnaliigijärjendite eelistatud järelkontekst %

Nimisõna		Tegusõna		Asesõna		Määrsõna	
DVG	91,07	DVX	72,73	IDJ	30,38	IDV	22,98
DVA	72,85	IPD	46,69				
IDN	69,39	IDP	46,08				
DVY	57,88	IDS	40,41				
DVP	43,51	IJD	35,46				
DVS	40,94	IID	34,26				
DKV	36,19	IDD	32,11				
AVD	31,17	ISD	31,65				
DVD	30,37	JID	26,05				
IDA	30,03	IAD	26,92				
PVD	26,35	JDI	24,74				
IVD	20,47	DPV	24,65				

Sealt alates on usaldus pikkamööda .. suurenenud; AVD<S, nt *Vedavateks jäid ikka tagarattad*; DVD<S, nt *Muidugi on ka kompromiss võimalik*; IDA<S, nt *No nii head aega ei osanud ma küll loota*.

b) Tegusõnaga: IPD<V, nt *No tema juba oskab lepingut lugeda*; IDP<V, nt *No kust te teate*; IDS<V, nt *Noh, kuidas eluke veereb?*; IJD<V, nt *Vat ja siis polnud muud kui minek*; IID<V, nt *Ah jaa, pisut lisa soola ja pipart*; IDD<V, nt *Vat nii siis teemegi*; ISD<V, nt *Oi jumal küll, kaduge minema*.

Nimisõna eelistatakse veel kahe ja tegusõna kolme trigrammi järel, kuid need juhud moodustavad kuni veerandi kasutusest: PVD<S, nt *Meie mõtlesime ka turvalisusele*; IVD<S, nt *No olgu siis asetäitja*; JID<V, nt *Kuid ennäe, nüüd hakkab iga arstile ilmumine raha maksma*; JDI<V, nt *Ning nüüd neh annab baltisakslane-nurjatu .. taas elumärki*; DPV<V, nt *Kas sa tahad minna matkale Karpaatidesse?*

Kahe trigrammi järel eelistatakse muid sõnaliike: IDJ-i järel asesõna, nt *No nii nagu teiegi*, ja IDV järel määrsõna, nt *No siiani pole veel vähemalt mulle arvet tulnud*.

Lingvistiline ülevaade. Lingvistilises vaates moodustavad üle poole lausealgulistest trigrammidest (kokku 15) hüüundid, nagu ISD (*Oi jumal küll!*). Enamasti algavad need hüüdsõnaga, kahel juhul ka sidesõnaga, nt JID: *Ja vat siis sööme vastlakukleid kah!* Üheksa järjendi esikomponent on määrsõna, nt DVD: *Samuti on veel võimalik osta .. .* Ühe järjendi algul esineb omadussõna (AVD: *Läbikukkunuks võib juba praegu lugeda plaani ..*) ja teisel asesõna (PVD: *Ma pole kunagi sellistel avamistel käinud*).

a) Hüüundid kuuluvad muutumatute sõnade klassi, tegemist on loendatava hulgaga. Nad võivad esineda iseseisva lausena (*No nii!*; *Oh kui raske!*) või selle süntaktiliselt sõltumatu osana (*Ahoi, kaugelt tulijad!*). Näiteid, kus hüüdsõna eraldatakse suhtlusvajadustest tulenevalt lause alguses komaga, on tavaliselt rohkem. Neid kasutatakse mh pöördumistes (*Hei, sina seal!*), tähelepanu köitmiseks (*Ohoo, keda ma näen!*) ja hinnanguteks (*Ja oh imet küll, asi toimib*). Koma ei kasutata lausealguliste sisutühjade hüüunditega, mis moodustavad neile järgneva(te) sõna(de)ga semantilise ja foneetilise terviku, nt *no nii*, *no ikka*, *no eks*, *no kas*, *no millal*, *no umbes*, *no näiteks*; *vaat/vat nii*, *vaat siis*, *vaat seal*; *ah ikka*.

Teine osa sisutühjadest hüüunditest paikneb pikema lause alguses, nt *Oi jumal küll*; *Oh kui tore*; *Säh sulle siis*; *No ja nüüd?*

Kolmas osa hüüundeid kuulub liitsete sidendite koosseisu, nagu *no isegi kui* või *no nii nagu*, mis lause alguses funktsioneerivad viiteseose vahendina, nt *No isegi kui see rahapada kuskil lebab – mida on meiesugustel vanadel inimestel enam nende kõlisevate müntidega peale hakata?*, ning sõnastust muutes võivad paikneda ka osalause alguses. Lausealgulisi hüüundeid *Ja vaat siis*, *Ja oh imet*, *Ja no muidugi* kasutatakse samuti viiteseose vahendina, rööpselt ka ekspressiivsete tähendusvarjundite edastamiseks.

b) Määrsõnaga algavad järjendid kinnitavad sõnajärjereeglit, mille kohaselt soovitatakse paigutada määrsõna lause algusse või lõppu. Siinse materjali põhjal on suurim tõenäosus paikneda lause alguses järjendil DVX koos väljend- ja ühendverbiga, nt *Seejuures unustati tähele panemata, et* .. ; *Seejuures unustati ära, et* .. . Oluline on märkida, et väljendverbi kuuluv käändsõna tuvastatakse analoogse kasutuse tõttu sageli, kuid mitte alati abimäärsõnana. Enam kui kolmandikul juhtudest esinevad lause algul järjendid DVY (*Eeldatavalt hakkab EVP hind .. taas langema*), DVS (*Tänavu saab riigilt põldude lupjamistoetust 330 põllumeest*) ja DVP (*Samas on Teie põlvkond käinud läbi peaaegu kõigist* ..). Väiksema levikuga on lause algul DVG, DVA, DVD, DPV ja DKV.

c) Omadus- ja asesõnaga algavad järjendid esinevad lause algul vähem kui kolmandikus keelenäidetes. Mustrid AVD ja PVD on tüüpilised V2-sõnajärje näited, kus öeldisele ja määrusele eelneb vastavalt omadussõna öeldistäite või määrusena (nt *Peamine on mitte langeda liialdustesse*) ning asesõna aluse või valdajamäärusena (nt *Mul on siin olnud häda pulsi allasaamisega*).

3.2. Lauselõpulised sõnaliigijärjendid

Statistiline ülevaade. Määrsõna sisaldavatest trigrammidest (kokku 303) on vaid viis sellist, mida eelistatakse lause lõpus (vt tabel 4). Nende esinemus koondkorpuses paigutub sarnasesse vahemikku lausealguliste trigrammidega (0,0001–0,833%).

Enam kui pooltel juhtudel esineb lause lõpus DPX, mida leidub korpuses väga vähesel määral. Ülejäänud nelja sõnaliigijärjendi esinemus selles positsioonis on poole väiksem, sh korpuses üsna levinud SDV-l ja DDV-l. Niisiis ilmneb sama seaduspärasus nagu lausealguliste trigrammidega: üleüldiselt suurema osakaaluga järjendeid kasutatakse lause lõpus väiksema tõenäosusega.

Tabel 4. Lauselõpulised sõnaliigijärjendid reastatuna lause lõpus esinemise tõenäosuse järgi

Jrk	Järjend	Osakaal koond-korpuses %	Osakaal lause lõpus %	Näide
1.	DPX	0,0001	54,72	Ei olnud *sinna midagi parata*.
2.	KDV	0,057	26,5	Nende kõigiga tuleb ehitamise või maa ostmise *üle läbi rääkida*.
3.	DNK	0,0042	25,0	Inimkatseid on tehtud *vaid ühe puhul*.
4.	SDV	0,833	23,04	Me ei ole *komisjoni veel moodustanud*.
5.	DDV	0,430	22,09	Suuliselt oleme neile *juba ära öelnud*.

Kuna DPX sisaldab abimäärsõna, siis kasutatakse selle ees regulaarselt tegusõna (vt tabel 5), nt *Savisaar on järgmine president, aga sind pisikest putukat ei pane varsti keegi tähele*. Teine pool DPX-i näidetest on sellised, kus DPX asub lause sees ja selle järel varieeruvad kuus sõnaliiki, millest suurem tõenäosus on tegusõnal (19,81%) ja sidesõnal (16,04%), nt *Eks ta ole isegi seda tähele pannud* .. ; *Varsti viiekuuseks saavad õde-venda panevad juba teineteist tähele ning vastavad teineteise kilgetele*.

Ülejäänud nelja trigrammi esinemus lause lõpus hõlmab leitud keelenäidetest umbes veerandi, mis tähendab, et suuremas osas lausetest paikneb mustri järel veel vähemalt üks sõna. Vahe varieeruvate sõnaliikide osakaalus on suhteliselt väike. KDV, SDV ja DDV järel on sagedamini tegusõna (vastavalt 19,11%, 20,26% ja 18,86%) või nimisõna (vastavalt 18,36%, 19,5% ja 22,09%), nt KDV<V: .. *raha selle eest välja ei käida*; SDV<V: *Seda seadused lihtsalt ei võimalda*;[14] DDV<S: *„Nad pakkusid kõige rohkem," põhjendab Holsting valikut*. DNK järel esineb lähedase osakaaluga nii tegusõnu (15,35%), määrsõnu (13,06%), omadussõnu (12,82%) kui ka nimisõnu (12,74%), nt DNK<D: *Seda on tavapärasest umbes kolmandiku võrra enam*.

Lauselõpuliste trigrammide ees, v.a tegusõnaühenditega seostuv DPX, kasutatakse kõige sagedamini nimisõna (vt tabel 5). KDV ees on regulaarselt kaassõna laiendav nimisõna (85,04%), nt .. *kõike saab täiuslikkuse poole edasi arendada.* Teiste trigrammide ees vahelduvad eri sõnaliigid. DNK ees võib lisaks nimisõnale (nt *Augusti lõpu seisuga kasvas klientide*

[14] *ei* kui tegusõna liitvormi osa märgendatakse tegusõnana.

arv vaid paarisaja võrra) esineda tegusõna (34,52%), harvem määrsõna (10,5%). Ka DDV ees on nimisõnaga (nt *Jutt on projektist, mille linnavalitsus esialgu välja töötas*) peaaegu sama sage tegusõna (24,5%), kasutatakse veel määr- (12,77%) ja asesõnu (11,19%) ning 13,38% juhtudest esineb DDV lause alguses. SDV ees leidub peale nimisõna (nt *Samal ajal on aga ka mitteaktiivsete noorte arv tunduvalt kasvanud*) ka tegu- (16,16%), omadus- (15,2%) ja asesõnu (14,25%).

Tabel 5. Lauselõpuliste sõnaliigijärjendite eelistatud eeskontekst %

Nimisõna		**Tegusõna**	
KDV	85,04	DPX	83,02
DNK	36,09		
SDV	29,36		
DDV	26,99		

Lingvistiline ülevaade. Lause lõpus eelistatud sõnaliigijärjendid on keeleliselt erinevad. See on ka mõistetav, kui silmas pidada, et lause alguses kasutatakse tavaliselt stereotüüpsema vormistusega sõnaliigijärjendeid kui lause lõpus. DPX-i kasutuses on leksikaalselt kinnistunud väljendverb *tähele panema* ja väljend *ei ole / pole midagi parata*, kuid vabalt võivad esineda ka muud väljend- ja ühendverbid. KDV ja DDV puhul kasutatakse ühendverbe, nt *tuleb maa ostmise üle läbi rääkida*; *Me püüame omavahel kokku leppida*. SDV puhul eelistatakse tegusõna pöördvormi, nt *Tele2 seevastu pole praktiliste töödega selles vallas veel alustanudki*.

KDV ja DNK on seotud kaassõna kasutamisega erinevates tingimustes. KDV ees paikneb reeglina nimisõna, nt *Sinna hakkavad kuuluma mehed, kes .. peavad esimese kutse peale kohale ilmuma*. DNK puhul aga järgneb kaassõna arvsõnale, moodustades leksikaalgrammatiliselt kinnistunud väljendeid, nagu *ligi kolmandiku võrra*, *vaid ühe puhul*, *eeldatavalt 26 kohta*, *valiti välja 939 hulgast/seast*, *kusagil 100 juures*, *vaid 10–15 ringis*, *juba kolme-nelja ajal*, *umbes 80-ndate keskel*, *siis 19-20 vahel*. Näitelausete põhjal võib öelda, et üldjuhul valitakse KDV ja DNK kasutamisel päritolult, tähenduselt ja abstraktsuse astmelt eristuvaid kaassõnu. KDV alguses on levinud koha- ja ajatähenduslikud kaassõnad (nt *üle*, *peal-peale*, *juures*, *poole*, *ajal*), järjest laiemat kasutust leidvad abstraktsed kaassõnad (vrd *laua peale* ja *kutse peale*) ning kaassõna funktsioonis kivinenud sõnavormid (*puhul*, *suhtes*, *kohaselt*, *järgi*, *jooksul*, *jaoks*, *tõttu*). DNK

puhul järgneb arvsõnale valik semantiliselt piiratud kaassõnu. Tegemist on umbmäärast aega, hulka või (aja)vahemikku (*ringis*, *keskel*, *ajal*, *vahel*, *juures*), määra (*võrra*), kohta (*hulgast/seast*) ja korduvust (*korral*) tähistavate väljenditega.

SDV ja DDV ees paiknevad sagedamini nimi- ja tegusõnad, kuid ka mitmed muud sõnaliigid. Nende kasutuses ilmnevaid tendentse tuleks analüüsida üksikasjalikumalt, mis pole aga siinse uurimuse eesmärk.

4. Veatuvastus sõnaliigijärjendite alusel

Hindamaks sõnaliigijärjendite leidja potentsiaali automaatses veatuvastuses, eraldasime testkorpusest ebatüüpilise kontekstiga trigrammid. Määrsõna sisaldavaid trigramme, mille ees- või järelkonteksti tõenäosus on alla 5%, leidus korpuses 599. Rühmitasime saadud järjendid selle järgi, kas need kattuvad mõne veamärgendusega. Ühe veakohaga võib kokku langeda mitu tri- või tetragrammi, samas võib üks järjend hõlmata mitut viga. Oleme arvesse võtnud nii täielikku kui ka osalist kattuvust. Täieliku kattuvuse korral sisaldab sõnaliigijärjend vigasena märgendatud lauseosa tervikuna. Osaline kattuvus tähendab, et vigane lauseosa algab enne või lõppeb pärast järjendit.

Lähtudes veaga kattuvate ebatüüpiliste järjendite osakaalust, on veatuvastuse täpsus 74,3%. Täpseim on A2-taseme tekstide veatuvastus, keeleoskustaseme tõustes suureneb väärpositiivsete tulemuste hulk (vt tabel 6). Lausestruktuur muutub kõrgematel tasemetel keerukamaks ja mitmekesisemaks ning kasutatakse enam harvaesinevaid sõnaliike, nagu kaassõnad ja lühendid, mis on trigrammide kontekstina vähetõenäolised (vt 4.3).

Tabel 6. Määrsõna sisaldavate ebatüüpiliste sõnaliigijärjendite kattuvus testkorpuse veamärgendusega

	A2	**B1**	**B2**	**C1**	**Kokku**
Ebatüüpilisi sõnaliigijärjendeid	85	108	190	216	599
Neist veaga kattuvad	73	85	146	141	445
Veatuvastuse täpsus %	85,9	78,7	76,8	65,3	74,3

Vigu hõlmavaid järjendeid liigitame selle alusel, kas tuvastatud viga mõjutab sõnaliikide järgnevust. Ligi poole leitud vigadest moodustavad sõnajärjevead, neist sagedaimad eiravad V2-sõnajärge lause või osalause alguses (vt tabel 7). Esineb ka liigse ja puuduoleva sõna, harvem sõnavaliku vigu. Lisaks võivad eksimused õigekirja ning kokku- ja lahkukirjutuse reeglite vastu tingida vea sõnaliigimärgenduses, mille tulemusena leitakse ebatüüpiline järjend.

Kui sõnajärjevigu tuvastatakse rohkem B1–C1-tasemel, siis ülejäänud veatüüpide osakaal hakkab pärast A2-taset vähenema. Teisalt muutuvad B2-tasemel sagedamaks juhud, kus esineb mitu veatüüpi korraga. Näiteks lauses *Kõrge rohkem mina kardan ~ Kõige rohkem kardan ma, et ..* on eksitud sõnajärjes ja sõnavalikul.

Eraldi loetleme vähetõenäolisi järjendeid, millega kattuvad vead ei seostu sõnaliikide järgnevusega. Neis järjendites tulevad kõige sagedamini esile puuduva koma, sõnavaliku- ja käändsõna vormivaliku vead. Kasutaja seisukohast ei ole tegemist tarbetu veatuvastusega, ent tuleb arvestada, et selliste juhuslike kattuvuste esinemus varieerub tekstiti.

Tabel 7. Veaga kattuvate sõnaliigijärjendite jaotumine veatüübi järgi

Veatüüp	A2	B1	B2	C1	Kokku
Eksimus V2-sõnajärje vastu	9 (12,3%)	34 (40%)	36 (24,7%)	43 (31,2%)	122 (27,4%)
Muu sõnajärjeviga	11 (15,1%)	16 (18,8%)	31 (21,2%)	27 (18,4%)	85 (19,1%)
Puuduv sõna	15 (20,5%)	7 (8,2%)	14 (9,6%)	7 (5,0%)	44 (9,9%)
Liigne sõna	9 (12,3%)	8 (9,4%)	5 (3,4%)	10 (7,1%)	32 (7,2%)
Sõnavalikuviga mõjutab sõnaliiki	6 (8,2%)	3 (3,5%)	3 (2,1%)	2 (1,4%)	19 (4,3%)
Õigekirja- või kokku-lahkukirjutuse viga mõjutab sõnaliigi-tuvastust	6 (8,2%)	6 (7,1%)	3 (2,1%)	4 (2,8%)	19 (4,3%)
Mitu viga korraga	3 (4,1%)	4 (4,7%)	12 (8,2%)	14 (9,9%)	32 (7,2%)
Viga ei seostu sõna-liikide järgnevusega	14 (19,2%)	7 (8,2%)	42 (28,8%)	34 (24,1%)	98 (22,0%)

Määrsõna sisaldavatele sõnaliigijärjenditele tugineva veatuvastuse täpsus sarnaneb tulemusega, mis on leitud kõiki vähetõenäolisi järjendeid arvestades (75,0%). Saagiseid võrreldes (11,6% *vs.* 39,2%) saab järeldada, et määrsõnaga järjendid toovad esile ligi kolmandiku kõigist leitud vigadest. Veatuvastuse täpsuse ja saagise arvutamisel oleme lähtunud iga lause puhul märgendusvariandist, mis sealsete ebatüüpiliste järjenditega enim kattub. Kui nt esimeses märgendusvariandis on järjendiga samas piirkonnas tähistatud sõnajärjeviga, ent teises pole vigu välja toodud, siis oleme arvestanud esimese variandiga. Juhul kui märgendusversioonid kattuvad rakenduse tähistatud võimalike veakohtadega võrdsel määral, oleme arvesse võtnud väiksema indeksiga varianti.

Järgnevalt anname ülevaate määrsõnaga sõnaliigijärjendites ilmnenud veajuhtumitest. Näitelausetes on poolpaksus kirjas tähistatud ebaloomulik sõnajärg, liigsed ja sobimatud sõnad. Puuduvad sõnad ja kirjavahemärgid on nurksulgudes. Vead, mis ei esinda kirjeldatavat veatüüpi, on tähistatud tärniga. Mõne veatuvastuse puhul võib vaielda, kas tegu on veaga. Veamärgenduses on tähistatud ka juhtumid, mis oleks vabamas keelekasutuses aktsepteeritavad, ent kirjalikus tekstis mõjuvad pigem ebaloomulikult.

4.1. Sõnajärjevead

V2-sõnajärg. Lause algul eksitakse V2-reegli vastu enamjaolt siis, kui lause algab määruse või määrusliku fraasiga. Harvem eksitakse öeldise paigutamisel lausetes, mis algavad või peaksid algama aluse või sihitisega. Lause alguse V2-sõnajärjest kõrvalekalded saab liigitada järgmiselt.

a) Sõnajärje määrus – öeldis – alus asemel kasutatakse järge määrus – alus – öeldis, nt *Praegu Anna on väga hea õpetaja ..* (D>SVD); *Teiselt poolt **tehnoloogia kiire** **areng annab** inimestele palju võimalusi* (^>PDS, P>DSA).

b) Määrus – öeldis – määrus(ed) asendatakse sõnajärjega määrus – määrus(ed) – öeldis, nt *Mul **praegu kodus** **ei ole** lemmikloomi ..* (^>PDS); *Kahjuks **Eestis on** praegu niisugune situatsioon ..* (D>SVD).

c) C1-tasemel on paaril juhul järjendi määrus – öeldis – öeldistäide asemel järjend määrus – öeldistäide – öeldis, nt *Teise ettekandja sõnul **kõige *vulnerabiilsemad ~ haavatavamad on** just lapsed ..* (N>SSD, A>VDS).

d) Järjendis alus – öeldis – määrus on vahetuses aluse-öeldise või öeldise-määruse asukoht, nt ***Elab ta*** *väga rikkalt* (^>VPD); *Nad* ***alati aitavad*** *raskel hetkel* .. (PDV<A).

Osalause algul esineb raskusi määrus(t)e, öeldise ja aluse järjestamisel. Neile võib eelneda ka rinnastav (*ja*, *ning*, *aga*) või alistav sidesõna (*et*, *kui*, *sest*). Eristuvad järgmised veajuhud.

a) Osalause algab sidususvahendina kasutatava määrsõnaga (*siis*, *seega*, *seepärast*, *veel*, *samuti*). Määrus – öeldis – alus/määrus on asendatud järjenditega, kus öeldisele eelnevad määrus(ed) ja alus või kaks määrust, nt *Ajakirjad on väga huvitavad naiste jaoks, peamiselt* ***need on*** *moest ja ilust* (K>DPV); .. *samuti* ***pargis puuduvad*** *isegi* **istukohad ~ istekohad* .. (D>SVD).

b) Sõnajärje sidend – määrus – öeldis asemel tarvitatakse järge sidend – määrus – alus. Osalause võib olla eelneva osalausega nii rinnastus- kui ka alistusseoses, vastavalt nt *Mina olen vana inimene ja minu jaoks* ***see on*** *väga raske* (K>PVD); *Tahan, et meie raamatukogus* ***kõik *jäätsid ~ jääks*** *nii, nagu praegu* (PVD<J).

c) Sidend – alus – öeldis asendatakse sõnajärjega sidend – alus – määrus, nt .. *olid keerulised ajad minu jaoks ning ma* ***juba kaotasin*** *lootust* .. (K>JPD). Alistava sidesõna puhul kasutatakse ka järjendeid sidend – määrus – alus ning sidend – alus – öeldistäide, nt *Ma valisin selle sellepärast, et* ***koolis ma lugesin*** *seda raamatut* (P>DJS<P); *Saan ka öelda, et nad* ***edukad on väga*** *nii töökohal kui ka isiklikus elus* (A>VDD).

d) Küsiva-siduva asesõnaga algavates kõrvallausetes vahetatakse öeldise ja määruse asukoht järjendis alus/öeldistäide – öeldis – määrus, nt .. *sellel põhjusel ei tööta ka digiseadmed, mis* ***omakorda viib*** *erinevate organisatsioonide töö *ajutisele peatumisele ~ ajutise peatumiseni* (PDV<A).

V2-sõnajärje vigadele osutab korduvalt 21 sõnaliigijärjendit, millest sagedamad on D>SVD (8), K>PVD (5), PVD<J (5), DVP<J (4) ja ^VDV (4). Järjend D>SVD viitab sõnajärjele määrus – alus – öeldis – määrus nii lause kui ka osalause algul, kus öeldis peaks alusele eelnema. K>PVD puhul järgnevad alus, öeldis ja määrus määruslikule kaassõnafraasile. Ka PVD<J tähistab määrusele järgnevat alust, öeldist ja määrust, mis esinevad koos rinnastava või alistava sidendiga osalausete piiril.

Järjendis DVP<J järgnevad määrusele või alusele määrus ja öeldis. Asesõna võib vastavalt olla kas alus või määrus ja kasutatakse alistavat sidesõna *et*, nt *Selles kollektiivis* ***mulle väga meeldib*** *see, et* .. ; *Probleem* ***aga *koosneb*** *sellest, et* ~ ***seisneb aga*** *selles, et* .. .

VDV lause algul hõlmab tegusõna pöördelist ja käändelist vormi. Tegu võib olla ahelverbi või tegusõna liitvormiga, samuti juhtumiga, kus alus on *da*-tegevusnimi. Määrsõna saab olla nii täistähenduslik kui ka abimäärsõna, nt ***On tihti*** *kuulda, et* .. ; ****Olid ~ Said läbi*** **rääkitud ~ räägitud mõned *variantid ~ variandid*.

Mõnede lausealguliste sõnajärjevigade puhul leiab sobiva paranduse järjendite seast, mida eelistatakse kasutada lause algul (vt 3.1). Näiteks tuleks DSV-le eelistada järjendit DVS (*Kahjuks Eestis on ~ on Eestis*). PVD võib paranduseks sobida, kui lause algul esineb PDV (*Nad alati aitavad ~ aitavad alati*), VPD (*Elab ta ~ Ta elab väga rikkalt*) ja PDS (*Teiselt poolt digiseadmete kasutamine tõi ~ tõi digiseadmete kasutamine*).

Muud sõnajärjevead. Ülejäänud sõnajärjevead puudutavad eelkõige olulisima info esitamist lause lõpus. Uut teavet kandva lauseosa, tüüpiliselt öeldistäite või sihitise asemel paigutatakse (osa)lauses viimaseks määrus(fraas), nt *Mina olin* ***nii rõõmus selle tõttu*** (VDA<P, DAP<K); *Tihti me jalutasime, käisime kinos, veetsime* ***vaba *aeg ~ aega koos*** *ja naersime* (ASD<J, DJV<$). Korduvalt on lause lõpus rõhusõna *ka*, mis peaks asuma rõhutatava sõna või fraasi ees, nt .. *ja peab austama* ***teiste inimeste õigust väärtuslikult *elad ~ elada ka*** (DVD<$).

Eraldi tõusevad esile laused, kus peaks määruse asemel viimasel kohal olema tegusõna käändevorm. Neis lausetes kasutatakse ahelverbe, *da*-tegevusnime sihitisena (öeldisega *tahtma*) või alusena (öeldisega *vaja olema*) ning mineviku kesksõnu öeldistäitena, nt .. *ja ma pean* ***olema peale operatsiooni temaga koos*** (K>SPD); *Ma tahan* ***palju *asja ~ asju teha seal*** (D>SVD, SVD<$); .. *meil on vaja* ***võtta kaks inimest juurde****, kuna töökoormus on suurenenud* (VDV<N); .. *need raamatud on* ***juba loetud *minuga ~ mul juba loetud*** (DVP<$).

Lisaks tarvitatakse (abi)määrsõna või määrust tegusõna asemel kaudküsilausete ja relatiivlausete lõpus, nt .. *mille pärast inimesed mõnikord* ***kolivad linnadesse*** .. (K>SDV); .. *mida enam* ***ei ole vaja****[,] ning* .. (VVD<J). Sobib ka V2-sõnajärg, ehkki seda ei ole veamärgenduses välja toodud. Sellised juhud on sagedamad C1-tasemel, kus pealause sõnajärjes eksitakse vähem.

Ilmneb eksimusi täiendifraasi komponentide ja põhisõna järjestamisel. Eestäiendi asemel on kasutatud järeltäiendit, nt ***Informatsioon, tulenev sealt**, kujundab nende* ~ ***Sealt tulenev informatsioon*** *kujundab nende maailmavaadet* .. (^>SAD, A>DVP). Täiendifraas võib paikneda põhisõna ümber, nt *Mulle meeldib* ****erinevaid muusikaid moodi*** ~ ***erinevat moodi muusika*** (ASD<$), või koosneda valesti järjestatud eestäienditest, nt lauses *Tean, et sina oled* ***väga pädev minu*** *kolleeg* (VDA<P), kus kuuluvust näitav isikuline asesõna peaks eelnema omadussõnalisele täiendile. Vealeidja toob esile ka raskusi pea- ja kõrvallause sidumisel. Näiteks ei järgne kõrvallause sõnale, mida ta laiendab: ***See on väga oluline veel****[,] missugune inimene sa oled* .. (VDA<D, A>DPS).

Korduvaid sõnaliigijärjendeid, mis eelkirjeldatud sõnajärjevigadele viitavad, on 16. Enim kasutatud järjendid on VDA<P (4) ja SVD<$ (4). VDA<P sisaldab tegusõna *olema*, määr- ja omadussõnast koosnevat öeldistäitefraasi ja sellele järgnevat asesõna, mis võib esineda eri funktsioonides (eestäiend, määrus või kaassõna laiend) ning peaks käima öeldistäite ees (nt *on väga ohtlik minu jaoks* ~ *minu jaoks väga ohtlik*). SVD eksimused lause lõpus tulenevad sellest, et viimaseks peaks asetama tegusõna. Asenduseks võib sobida eelistatult lause lõpus esile tulnud järjend SDV (vt 3.2; nt *vaatan, milline ilm on homme* ~ *homme on*), kuid sõltuvalt kontekstist ka DSV, nt hulgafraasi korral (*tahan palju *asja* ~ *asju teha seal* ~ *seal palju asju teha*).

4.2. Muud sõnaliikide järgnevust mõjutavad tegurid

Puuduvad ja liigsed sõnad. Valdavalt puudub lausest tegusõna *olema* olevikuvorm, mida võib pidada vene emakeele mõjuks. A2-tasemel ilmneb see viga sagedamini kogeja-omajalausetes, nt **Ma* ~ *Mul [on] väga hea meel* (^>PDA), kuid ka normaallausetes, nt *Seal [on] väga soe* (^>DDA). Neil juhtudel sobiks paranduseks lausealguline PVD (vt 3.1). Kõrgematel tasemetel puudub *olema* enim lausetest, kus seda laiendab öeldistäide või seisundimäärsõna, nt **Kui* ~ *Kuna laps [on] praegu haige, siis tahan temaga kodus olla* (J>SDA<D).

Lausetest võivad välja jääda muudki tegusõnad või ühendverbi kuuluvad abimäärsõnad, nt *See kodumasin [läks] katki eile* (^>PSD); *On vaja [välja] selgitada[,] milline on *peasõna ristsõnas* ~ *ristsõna lahendus* .. (^>VDV). Samuti leidub puuduva sidesõna näiteid, nt *Kujutan ette, [et]*

ma tulen koju .. (^>VDP), ja juhtumeid, kus nimi- või asesõna väljajätt muudab lause raskesti arusaadavaks, nt *[Temaga/Koeraga] ei ole vaja *ulutada ~ jalutada õues* (^>VVD).

Liigsed sõnad on peamiselt funktsioonisõnad. Tarbetult on kasutatud kaassõnu *poolt*, *puhul*, *peal* ja *juures*, nt *Osalesin hiljuti Päästeameti **poolt** korraldatud teabepäeval* (VDS<K); *Olukord kiirtee **peal** ~ kiirteel on veel kurvem* (K>VDA). Kui *poolt* on sisult ülearune, siis teiste kaassõnade asemel sobib kasutada alaleütleva käände vormi (nt *teenuste **puhul** ~ teenustel ei ole head kvaliteeti*, *kontserdi **peal** Dima Bilan **juures** ~ Dima Bilani kontserdil*).

Mõnikord tekib liiasus määrsõnakorduse (*seal*, *ka*, *ei*) tõttu, nt *.. paljud uued elanikud ei osale kohalikus elus **ei*** (ASD<$); *.. mitte ainult infot, mis tuleb ümbrusest[,] *kui ~ vaid ka seda, mis *tuleneb ~ tuleb **ka** endalt* (VDP<$). Mitmel juhul on liigsed tegusõna *olema*, määrsõna *siis* ja sidesõna *et*, nt *Ma loodan[,] et **on** näeme veel* (J>VVD, VVD<$); *Ning **et** kui midagi siin teid huvitab ..* (J>JPD).

Sõnavalik ja õigekiri. Üksteise asemel kasutatakse kõlalt ja/või tähenduselt lähedasi omadus- ja määrsõnu ning ase- ja määrsõnu. Sagedamini tuleks määrsõna asendada omadussõnaga (*kiiresti ~ kiire*, *hästi ~ hea*, *vähem ~ väiksem*), nt *.. austamine on kõige vähem ~ väiksem tegu[,] mida me saame teha ..* (D**D**S<P). Esineb ka vastupidine näide: *.. seepärast on *väga vaja **hea** mõtelda, kas võta või ei kassi korterisse ~ vaja väga **hästi** mõtelda, kas võtta kassi korterisse* (**A**>VDV).

Mõlemat pidi on omavahel vahetatud samuti määr- ja asesõnu, nt *Ma puhkasin *hottelis, **kus** asutas mere juures ~ hotellis, **mis** asus mere ääres* (**D**VS<K); *Käisime koos *Astrikeskusse **sind** tihti olevad kontserti ~ Astri keskuses, **seal** on tihti kontserdid* (K>S**P**D ~ D>S**P**D). Viimases näites võib olla silmas peetud määrsõna *siin*. Lisaks on asemäärsõna *kus* asemel tarvitatud sidesõna *kui*: *.. koeraga on võimalik *jooksta metsas ~ joosta metsas, kui ~ kus on päris puhas õhk* (**J**>VDA ~ **D**>VDA).

Osa õigekirjaveaga sõnu on homonüümsed muud liiki sõnaga. Sellised juhtumid on sõnaliigituvastuse jaoks sarnased sõnavalikuvigadega, nt *Paljud arvavad, et raamatupidaja töö on kergem kui teised tööd, aga see poole ~ pole nii* (P**K**D<$ ~ P**V**D<$); *Seal on vaga ~ väga ranged *ohtusnõued ~ ohutusnõuded ..* (DV**A**<A ~ DV**D**<A); *Naga ~ Nagu ma juba *kirjutas ~ kirjutasin ..* (^>**S**PD ~ ^>**J**PD). Lisaks on õigekirjaveaga

ase-, määr- ja tegusõnu määratud nimisõnaks (nt *mule ~ mulle*, *ümbes ~ umbes*, *bronerime ~ broneerime*) ning nimisõnu omadussõnaks (nt *maister ~ meister*, *promenad ~ promenaad*).

Kokku- ja lahkukirjutuse vead on seotud liitsõnade lahkukirjutuse ja ühendverbide kokkukirjutusega, nt *Või .. on parem *kõike ~ kõik otsekohe **prüügi kasti ~ prügikasti** panna?* (P>DS**S**); *Päästeameti töötajad soovitasid .. mitte juua rohkem kui organismile **ettenähtud ~ ette nähtud*** (DJS<**A**).

4.3. Tarbetud veatuvastused

Vealeidja toob esile ka sõnaliigijärjendeid, mis väiksest esinemistõenäosusest hoolimata veamärgendusega ei kattu. Neid tasub lähemalt vaadelda, et parandada veatuvastuse täpsust. Siinsete testandmete põhjal saab välja tuua neli sagedamini korduvat määrsõnaga järjendit, millega kaasnevad väärtuvastused: VVD<J (7), PVD<J (6), PVD<$ ja SVD<$. Kuna need võivad teatud tingimustel osutada sõnajärjeveale, on oluline arvesse võtta ka eelnevat konteksti.

Järjendit VVD<J kasutatakse a) juhul, kui soovitakse rõhutada määrust, mis paigutatakse tegusõna käändelise vormi asemel (osa)lause lõppu, nt *Nad pole märgitud selgelt ja õigesti*; *Me võiksime minna koos, kuna* .. ; b) eitavas lauses, eelistades V2-sõnajärge, nt *Minu ema ei tea veel, et ma tahan koera*. Vähem loomulik on see, kui mustrile eelneb aluse asemel määrus, nt *leidsin .. raamatuid ja ajakirju, mida **enam ei ole vaja**[,] ning* .. (vt ka 4.1).

Ka PVD (osa)lause lõpus eirab V2-sõnajärge, kui sellele eelneb määrus (vrd *Me jalutasime palju ja ujusime* ning *Kogu aeg **midagi läheb** katki ja* ..). Kui järjend sisaldab rõhusõna, võiks eelistada tegusõna lõpupositsioonis (nt *igaühel **juhtub ju***). SVD mõjub ebaloomulikult, hõlmates sihitist ja *da*-tegevusnime, mis laiendavad tegusõnu *tahtma* ja *saama* (vrd *Kontsert meeldis väga!* ja *Tahame järgmisel kuul ***teatri ettendus ~ teatrietendust vaadata ka***).

Tulemusi võib mõjutada õppijakeele spetsiifika, mh laialdasem asesõnakasutus, mis võib olla tingitud kirjutamisülesannete sisust ja tekstiliigist, nt kirjades kasutatakse rohkelt isikulisi asesõnu. Väärtuvastustega seotud sõnaliigijärjendites on asesõnad suhteliselt sagedad: 104 erinevast järjendist 12 esineb lause algul ja 14 lause lõpus, asesõna sisaldavad vastavalt 5 ja 6 mustrit. Ebatüüpiliste eeskontekstide (43) seas on asesõnad

side- ja kaassõnade (13) järel teisel kohal (9), ebatüüpiliste järelkontekstide (35) seas side- (12) ja arvsõnade (7) järel kolmandal kohal (6).

Kaas- ja arvsõnade, omastavalise täiendi ja lühendite esinemus trigrammide ees/järel ei ületa üldiselt 5% piiri. Näiteks kaassõna keskmine tõenäosus esineda trigrammi eeskontekstina on 2,7% ning 93,6% trigrammide puhul on tõenäosus alla 5%. Samas ei tasu neid kontekste kõrvale jätta, sest need aitavad mõningaid vigu tõhusalt tuvastada. Mõistlik oleks rakendada väiksemat piirmäära ja teha kindlaks veatuvastuse seisukohast kasulikumad järjendid.

5. Kokkuvõte

Artiklis esitatud uurimistulemused ilmestavad sõnaliigijärjendite ja nende konteksti potentsiaali nii keelekasutuse seaduspärasuste määratlemisel kui ka ebatüüpilise keelekasutuse tuvastamisel. Määrsõna sisaldavate järjendite analüüs näitas järgmist.

1. Lause alguses ja lõpus eelistatud trigrammide esinemus koondkorpuses paigutub sarnasesse vahemikku (alla 1%). Suurema osakaaluga trigrammidel on väiksem tõenäosus nendes positsioonides asuda.

2. Lause alguses eelistatakse stereotüüpse leksikaalgrammatilise vormistusega hüüundeid (harvem väljend- ja ühendverbe). Lause lõpus paiknevad sõnaliigijärjendid on keeleliselt mitmekesisemad, ulatudes ebamäärast hulka või mingit ajalist piiri tähistavatest väljenditest väljend- ja ühendverbide ning kaassõnafraasideni.

3. Määrsõnaga seotud sõnajärjevead kipuvad esinema lause alguses või lõpus, seega saab neid mõnel juhul parandada lausealgulise või -lõpulisena eelistatud sõnaliigijärjendite toel. Tõenäolisemalt sobivad paranduseks järjendid, mille osakaal koondkorpuses on võrdlemisi suur ja mis küll esinevad sageli lause alguses/lõpus, kuid mitte eelistatud variandina. Edaspidi, kui veatuvastuse kõrvale lisandub paranduste pakkumise funktsioon, on oluline ka neile järjenditele tähelepanu pöörata.

4. Keeleõppijate tekstides õnnestus esile tuua eelkõige V2-sõnajärje vigu. Tegu on peamiselt juhtumitega, kus alus ja määrus või mitu määrust paiknevad enne öeldist. Sõnajärjevead kaasnevad keerukama lausestruktuuriga, olles tunduvalt sagedamad B1–C1-taseme tekstides. A2-tasemel on sama sage viga puuduvad ja liigsed sõnad. Kuna sõnaliigijärjend, mis aitab vigu avastada, võib põhjustada ka tarbetuid veatuvastusi, on

oluline arvestada enama kontekstiga, nt ebatüüpilise järelkonteksti puhul ka eeskontekstiga.

Koondkorpuse sõnaliigi n-grammide analüüsi jätkates on plaanis kirjeldada lause sees esinevaid kontekste – eriti neid, mis kasutusel osalausete piiril, kus leidub sagedasti sõnajärjevigu. Kasulik oleks sõnaliigijärjendi ees- ja järelkonteksti vaadata koos, mitte eraldi. Veatuvastajat arendades tasub kirjakeele seaduspärade kõrval analüüsida õppija mittestandardset keelekasutust ning määrata, milliste järjendite alusel ja mis tingimustel saab selles kõige tõhusamalt vigu tuvastada.

Praeguseks üle 74% ulatuv veatuvastuse täpsus on suhteliselt hea tulemus, kuid õppijakeeles on arvestatav juhuslike kattuvuste osakaal. Veatuvastaja tulemuslikkust on kavas hinnata ka emakeeleõppija testkorpusega, millele on lisatud sama skeemi järgiv veamärgendus[15].

Kirjandus

Alam, Jahangir Md., Naushad UzZaman, Mumit Khan 2007. N-gram based statistical grammar checker for Bangla and English. – Proceedings of 9th International Conference on Computer and Information Technology, 3–6.

Allkivi-Metsoja, Kais, Jaagup Kippar 2023. Spelling correction for Estonian learner language. – Proceedings of the 24th Nordic Conference on Computational Linguistics, 782–788.

Aulamo, Mikko 2019. Using POS n-grams to detect grammatical errors in Finnish text. Magistritöö. Helsingi Ülikool.

Brett, David, Antonio Pinna 2015. Patterns, fixedness and variability: using PoS-grams to find phraseologies in the language of travel journalism. – Procedia – Social and Behavioral Sciences 198 (2015), 52–57. https://doi.org/10.1016/j.sbspro.2015.07.418.

Bryant, Christopher, Mariano Felice, Ted Briscoe 2017. Automatic annotation and evaluation of error types for grammatical error correction. – Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 793–805. https://doi.org/10.18653/v1/P17-1074.

Cappelle, Bert, Natalia Grabar 2016. Towards an n-grammar of English. – Applied Construction Grammar. Ed. by Sabine De Knop, Gaëtanelle Gilquin. De Gruyter Mouton, 271–302. https://doi.org/10.1515/9783110458268-011.

De Cock, Sylvie, Sylviane Granger 2021. Stance in press releases versus business news: A lexical bundle approach. – Text and Talk 41 (5–6), 691–713. https://doi.org/10.1515/text-2020-0040.

[15] https://github.com/TartuNLP/estgec/tree/main/Tartu_L1_corpus.

Eslon, Pille, Kais Allkivi-Metsoja 2018. Teksti keelekasutusmustrid ja lingvistiline klasteranalüüs. – Lähivõrdlusi 28. Lähivertailuja 28. Peatoim. Annekatrin Kaivapalu. Tallinn: Eesti Rakenduslingvistika Ühing, 21–46. http://dx.doi.org/10.5128/LV28.01.

EstGEC-L2 2023 = Estonian L2 Grammatical Error Correction Corpus (EstGEC-L2). Github. https://github.com/tlu-dt-nlp/EstGEC-L2-Corpus.

EstSpacy 2021 = SpaCy pipelines for Estonian language. Github. https://github.com/EstSyntax/EstSpaCy.

Jackendoff, Ray 2017. In defense of theory. – Cognitive Science 41 (S2), 185–212. https://doi.org/10.1111/cogs.12324.

Kapusta, Jozef, Martin Drlik, Michal Munk 2021. Using of n-grams from morphological tags for fake news classification. – PeerJ Comput. Sci. 7 (624). https://doi.org/10.7717/peerj-cs.624.

Luhtaru, Agnes, Mark Fišel, Elizaveta Korotkova 2024. No error left behind: Multilingual grammatical error correction with pre-trained translation models. – Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, 1209–1222.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning 2020. Stanza: A Python natural language processing toolkit for many human languages. – Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14.

Sirts, Kairit, Kairit Peekman 2020. Evaluating sentence segmentation and word tokenization systems on Estonian web texts. – Human Language Technologies – The Baltic Perspective, 174–181. https://doi.org/10.3233/FAIA200620.

Sõnaliigijärjendite leidja 2024. Github. https://github.com/tlu-dt-nlp/POSgram-contexts/blob/main/posgram_finder_demo_et.ipynb.

Sõnaliigijärjenditel põhinev veatuvastus 2024. Github. https://github.com/tlu-dt-nlp/POSgram-errors/blob/main/error_finder_demo_et.ipynb.

Wu, Jian-cheng, Jim Chang, Jason S. Chang 2013. Correcting serial grammatical errors based on n-grams and syntax. – Computational Linguistics and Chinese Language Processing 18 (4), 31–44.

Detecting regularities and errors based on adverb-containing POS-grams

KAIS ALLKIVI-METSOJA,
PILLE ESLON, JAAGUP KIPPAR

The article introduces a software tool that allows us to detect regularities and errors in Estonian language texts, based on the usage contexts of POS-grams. It converts each sentence to a POS string and extracts trigrams, i.e., three-word sequences. Then, it calculates the probabilities of various preceding and subsequent contexts, which can either be a certain POS, or the beginning or the end of a sentence. Error detection relies on the comparison with a statistical language model.

In this paper, we focus on the contexts of adverb-containing POS-grams, which are prone to word order errors. Our aim is two-fold: 1) using the Estonian Reference Corpus, we build a language model and analyse it to describe the POS-grams that are preferably used in the context of sentence onset or ending; 2) we evaluate the error detection performance of the tool on the EstGEC-L2 test corpus, consisting of error-annotated sentences from second language learner writings. The cut-off value for defining rare contexts is set to 5%.

We find that the POS-grams commonly used in sentence onsets are lexico-grammatically more stereotypical, while those preferred at the end of a sentence show more variation. POS-gram analysis also proves to be useful in pointing out word order errors, unnecessary and missing words, occasionally word choice and spelling errors (if POS detection is affected). Most frequently, the detected errors violate the V2 word order at the beginning of a sentence/clause. Other word order errors occur mainly at the sentence/clause ending.

Keywords: grammatical error detection, natural language processing, morpho-syntax, usage-based approach, n-grams

Kais Allkivi-Metsoja
digitehnoloogiate instituut
Tallinna Ülikool
Narva mnt 25
10120 Tallinn
kais@tlu.ee

Pille Eslon
digitehnoloogiate instituut
Tallinna Ülikool
Narva mnt 25
10120 Tallinn
peslon@tlu.ee

Jaagup Kippar
digitehnoloogiate instituut
Tallinna Ülikool
Narva mnt 25
10120 Tallinn
jaagup@tlu.ee