Pille Eslon, Jaagup Kippar
(Tallinna Ülikool, Digitehnoloogiate instituut)

Keelekasutusmustrid & kontekstid

ERÜ 20. kevadkonverents *Keel ja keelekasutajad*
27.-28. aprill, Tallinn

HTM projekt TKA21200 *Eestikeelse teksti automaatkorrektuur*
TLÜ uuringufondi projekt TF2621 *ELLE: eesti keele õpet toetav keeletehnoloogiline ressurss*

- **Vajadus**: arendada automaatse keeletöötluse vahend, mis tuleks toime nii emakeelse inimese kui ka eesti keele õppija kirjutatud tekstide keelekasutuse analüüsiga, annaks tagasisidet sõnade, vormide, tüüpiliste leksikaalsete koosluste ning õigekirja osas, pakuks paremaid sõnastusi ja viitaks reegli(te)le.
- **Mõte**:
  - leida, süstematiseerida ja lingvistiliselt analüüsida eri keelekasutusvariantides esile tulnud keelekasutusmustreid
  - teha kindlaks mustreid tüüpiliselt ümbritsevad sõnaliigid ja sõnaliigikooslused tekstis;
  - treenida vahend, mis oskaks neid tüüpilisi sõnaliigijärjendeid tuvastada ja eristada eesti keelele mitteomastest või harvaesinevatest sõnaliigijärjenditest

Keelekasutusmuster

- **Keelekasutusmustrid** (sõnaliigijärjendid, morfosüntaktilised struktuurid)
  - mustril on **püsikindel struktuur**, nt DDD (*kahjuks ka palju*, *Nagu eelpool juba*)
  - mustri komponentidel on **tüüpiline vormistus ja tüüpilised tekstifunktsioonid**, nt DDJ alusel moodustatakse ja kasutatakse vastandavat paarissidendit *ei … ega* (*ei rohkem ega vähem*), sünonüümsed struktuurid on DNJ ja DPJ (*ei üks ega teine ~ ei see ega teine*)
  - mustri **leksikaalgrammatilisel varieeruvusel on piirid**, nt VDD alusel moodustuvad ja kasutatakse ühendverbe (*käib kaasas ka* > kaasneb); hoogsat tegutsemist tähistavaid ütlusi (*sammume/liigume/põrutame kindlalt edasi*); piltlikke väljendeid ([ei] *vea enam välja*)
- Keelekasutusmustrid on samalaadsed muude mitmesõnaliste kooslustega, mis esinevad tekstis regulaarselt
  - sarjadena (Sahkai 2006, 2008, 2011; Muischnek ja Sahkai 2010)
  - ütluste ja lauseahelatena, kuuludes kinnistunud sõnakasutusega lauseliste kõnevormelite ja jätkuga lausete alla (Õim & Õim 2019)
  - fraasidena, millest poole moodustavad vormelid: leksikaalsed kimbud, keeletükid, mitmesõnalised väljendid või keeleüksused, mis eraldatud tekstist erinevate tekstitöötlusvahendite abil (Evert 2005; Muischnek 2006; Muischnek & Kaalep 2010; Schmitt 2010; Chen & Baker 2010; Martinez & Schmitt 2012; Juknevičienė 2013; Selivan 2018; De Cock & Granger 2021)

Uurimuse lähted ja vahendid

- **Kasutuspõhine uurimus** (Biber 2009 –> corpus driven approach) > sõnaliigijärjendid kui kasutuspõhised sõnajärjemallid
  - TÄHENDUS<=>TEKST teooria: lekseemide süntaktiliste ja leksikaalsete piirangute kirjeldamine pindmisel ja süvatasandil – alus lekseemide kombineerumisreeglite sõnastamisel (vt Mel'čuk 1995: 5–6)
  - Mustri struktuuri kui keelekasutuses korduvalt esinevasse sõnaliigijärjendisse on kodeeritud selle komponentide kooskasutuse fonoloogilised, süntaktilised ja kontseptuaalsed reeglid, mis aktiveeruvad keelekasutuses semantika-grammatika piirimail (ajendatuna Jackendoff 2017)
- **Keelekasutusmustrite eraldamine tekstidest**:
  - andmekaeve põhimõttel töötav programm Klastrileidja (Matsak jt 2010; Ots 2012) → arendus Mustrileidja (Liiva 2022)
  - vahendi loomine, mis eristab tüüpiliste sõnaliigijärjendite kujunemist mustrist vasakul ja paremal (arvestatakse esimene + teine sõnaliik mustri ees ja järel, mis moodustavad sageli/harva esinevaid sõnaliigikooslusi)

Valmidus

- 3-komponentsete keelekasutusmustrite lingvistilised kirjeldused
  - 3 komponenti – järjendi tavapärane pikkus keelekasutuse automaatse analüüsi puhul (nt De Cock & Granger 2021)
- mustrid leitud publitsistlikest arvamuslugudest, 1890ndate ja 1990ndate ilukirjandusest ning eesti õppijakeelest (nt Trainis 2017; Trainis & Allkivi 2014; Eslon 2022, 2014 a, b; Allkivi-Metsoja 2016)
- koostatud mustrite hierarhiline klassifikatsioon, läbi viidud süsteemne lingvistiline analüüs, kirjeldatud saadud tulemusi, sõnastatud keelekasutuse seaduspärasusi
- ettekandes kasutame andmeid **adverbi sisaldavate mustrite** kohta, mis leitud EVKK publitsistlike arvamuslugude referentskorpusest ja eesti keele koondkorpusest

Näide 1: mustri lingvistiline analüüs

- Järjend DDD – publitsistlikes arvamuslugudes keskmiselt produktiivne muster nii verbist vasakul kui ka paremal
- 1. komponent
  - tavaliselt määrus (aja-, põhjus- ja kaasnevusmäärus), nt *nüüd* (*nüüd ju ikkagi*), *praegu* (*praegu veidi valepidi*), *seejuures* (*seejuures mitte kunagi*), *miks* (*miks mitte kohe*)
  - harvem rõhusõna (*küll veidi pealt* vaatavad) või subjektiivmodaalne hinnangusõna, nt *paraku* (*paraku väga palju*), *kindlasti* (*kindlasti veel edasi* minna), *vaja* (*vaja palju rohkem*)
  - võib kuuluda tugiverbikonstruktsiooni, nt on *nõus solidaarselt koos* töötama

- 2. komponent
  - tavaliselt rõhusõna (nt *just*, *nimelt*, *kunagi*, *veel*, *ikka*, *nii*, *ka*)
    - moodustab sõnakooslusi eelneva (*vaja just* *üle* lugeda, *just nimelt* *praegu*, ei ole/pole *mitte kunagi* *vaja*) ja järgneva adverbiga (*mitte just kõige* teravam pliiats, tehti *lahti just nimelt* täna, *siiski veel vaid* unistus, *seal ikka nii* kesiselt, *pealegi nii ju* saab lihtsamalt, *Eks ka nii* võib)
  - niisugused adverbikooslused koondavad lugeja tähelepanu kõnesolevale, kujundades arvamust millegi olulisusest või vajalikkusest/mittevajalikkusest

- 3. komponent
  - adverb võib olla nii määruse, rõhusõna kui ka verbipartikli funktsioonis
  - verbipartikli korral võib põhiverb jääda
    - mustrist paremale (*väga põhjalikult läbi* uuritud)
    - või vasakule (on uuritud *väga põhjalikult läbi*)
- Mida saab järeldada mustri DDD lingvistilise analüüsi põhjal?
  - Adverbide funktsioonid varieeruvad olenevalt kontekstist
  - DDD esineb koos verbi sisaldavate predikatiivsete üksustega

- Mustrit DDD ümbritsevad
  - mineviku liitajavormide komponendid, nt on *seal ikka nii* olnud/elatud;
  - tugiverbikonstruktsiooni komponendid, nt on *kõikjal veel selgemalt* kirjas;
  - koopula ja predikatiiv, nt oli *mitte just päris* korrektne;
  - üldeitus, nt pole *praegu veel täpselt* teada;
  - ühendverbi komponendid – põhiverb võib paikneda
    - mustrist paremal (*väga põhjalikult läbi* uuritud)
    - või vasakul (on uuritud *väga põhjalikult läbi*)
- **Küsimus**: **millist sõnaliigikonteksti DDD kasutamisel statistiliselt eelistatakse?**

Näide 2: publitsistlikud arvamuslood (109 326 sõnet) – koondkorpus (~ 181 miljonit sõnet)

- Publitsistlike arvamuslugude statistikast nähtub:
  - DDD ees ja järel eelistatakse verbi – vastavalt 41% ja 30% kõigist kasutustest, ülejäänud sõnaliikide osakaal selles positsioonis on väiksem

J	P	D	S	V	DDD	V	S	D	A	P	J
5%	13%	16%	23%	41%		30%	20%	17%	16%	9%	5%

- Koondkorpuse statistikast nähtub (arvestatud ka lause alguse ja lõpu tähisega):
  - mustrist vasakul on samuti sagedam esinemus verbil – 39.59%, varieeruvad substantiiv (23.42%), adverb (11.12%), pronoomen (9.18%), konjunktsioon (7.29%), positsioon lause alguses (5.99%), adpositsioon (1.67%), adjektiiv (1.16%), numeraal (0.28%) jm
  - mustrist paremal samuti verb (27.3%), varieeruvad substantiiv (18.33%), adjektiiv (11.87%), adverb (11.12%), konjunktsioon (11.05%), positsioon lause lõpus (10.75%), pronoomen (5.81%), numeraal (2.71%), adpositsioon (0.55%), abreviatuur (0.32%) jm
- **Milliseid sõnaliigikooslusi mustrist vasakul ja paremal nende sõnaliikide varieerudes moodustub?**

Näide 3: võimalikud sõnaliigikooslused DDD ees ja järel publitsistlikes arvamuslugudes

- **Verb DDD ees (41%)**
  - subst
  - pron
  - verb
  - adpositsioon

Substantiiv DDD ees (23%)

- verb
- pron, subst
- adj
- adv, konj

Adverb DDD ees (16%)

- verb
- pron, subst
- adv, konj

- **Verb DDD järel (30%)**
  - verb
  - subst
  - konj
  - pron
  - adv

Substantiiv DDD järel (20%)

- subst
- verb
- konj
- adv
- adpos, pron

Adverb DDD järel (17%)

- verb
- subst
- adj, adv, konj

Mustrile eelneva/järgneva konteksti tuvastamine publitsistlike arvamuslugude alusel adverbi sisaldavate mustrite kaupa: tehnoloogia

- Uuritavast tekstist parseriväljund (vislcg3)
- Laused ja sõnad koos metaandmetega mällu
- Vajalike metaandmete (sõnaliik, morfosüntaksi osa) eraldamiseks käsklused
- Võimalus ühendada laused, eemaldada kirjavahemärgid
- Otsimisfunktsioon, kus päring parameetrina, nt ..DDV ja DDV..
- Statistika, näidete esitamine

Arvutustulemuste esitamine publitsistlike arvamuslugude adverbi sisaldavate mustrirühmade kaupa: mustrirühmad

Morfosüntaktiliste mallide arv	Eristav tunnus	Mustrid	Mustrite arv
Üks mall	Morfosüntaktiline stereotüüp	SKD, DAV, DVA, DAJ, DSK, DPV, VJD, VDJ, PSD	9
Mitu malli	Üks lingvistiline tunnus	SJD, SDD, DDS, DSS, DAS, DGS, DDA, DSJ, DVS, JDS, JVD, JSD, JDA, VDS, ASD, PDD	16
	Aheltunnus	VDD, VVD, SVD, DVV, DDV, DVD, DSV	7
	Komplekstunnus	DPS, VSD, VDA, PVD	4
	Formaaltunnus	DDD, DJD, DDJ, JDD, JDJ	5

```
mustrid={

"morfosyhtaktiline_stereotyyp":["SKD",
"DAV", "DVA", "DAJ", "DSK", "DPV",
"VJD", "VDJ", "PSD"],

  "yks_lingvistiline_tunnus":["SJD",
"SDD", "DDS", "DSS", "DAS", "DGS",
"DDA", "DSJ", "DVS", "JDS", "JVD",
"JSD", "JDA", "VDS", "ASD", "PDD"],

  "aheltunnus":["VDD", "VVD", "SVD",
"DVV", "DDV", "DVD", "DSV"],

  "komplekstunnus":["DPS", "VSD",
"VDA", "PVD"],

  "formaaltunnus":["DDD", "DJD",
"DDJ", "JDD", "JDJ"],

}
```

Kontekstid ja grupid

Kõigepealt sobivad vasted välja

```
vasted=otsi(lausepuu, ".."+jada)
```

Siis soovitud read esitamiseks ühendada

```
  ryhmad=ryhmita(vasted, lambda kirje: "-".join(
    [sonaliik(rida) for rida in
kirje[:2]]+[morfosynt(rida) for rida in kirje[2:]]))
```

edasi juba soovitud kujul väljatrükk

```
  tryki_mitu(ryhmad, lambda ryhm:
[sonaliik(ryhm[0][0])+sonaliik(ryhm[0][1]),
     kogus(ryhm), …
```

Mustrile eelneva/järgneva konteksti tuvastamine koondkorpuse alusel adverbi sisaldavate mustrite kaupa

VVD vasakkonteksti näide

- V;5;0.01;V aux neg // @NEG;V mod indic pres ps neg // @FCV;V main inf // @IMV;D // @ADVL;ei saa koguda vaid,ei või olla nii,ei saagi areneda ei,ei saa olla vaid,ei tohi unustada ka
- S;5;0.01;S com sg part // @OBJ;V mod indic pres ps3 sg ps af // @FCV;V main inf // @IMV;D // @ADVL;tähelepanu võib ennustada ka,Nihet võib täheldada ka,nihet võib näha ka,seadust saab muuta ainult,toetust saab jagada mehhaaniliselt

Mahukama võrdlusandmestiku loomine koondkorpuse näitel (181 miljonit sõnet)

- Piisavalt kiireks märgendajaks osutus Stanza koos protsessoritega tokenize, pos – üks arvuti saab nädalaga hakkama
- Tekstid lausete kaupa sõnaliigijadaks, juurde lause alguse ja lõpu märk
- Näide: Tallinna linn algatab Paldiski maantee ääres Hotell Tallinna kõrval asuva suure vundamendiaugu ja tühermaa detailplaneeringu koostamise, ehitustööde alustamist takistavad aga ala segased omandisuhted.
- ^SSVSSKSSKAASJSSSZSSVDSASZ$
- Edasi sealt juba kolmikud koos vajalike eelnevate ja järgnevate sõnaliikidega välja.

SVD,2456405,1.6095%
S,778937,31.71%
^,420644,17.12%
A,368309,14.99%
J,280102,11.4%
P,247095,10.06%
V,144818,5.9%
D,94400,3.84%
N,66468,2.71%
K,25441,1.04%
Y,23226,0.95%
G,5549,0.23%
I,785,0.03%
X,614,0.02%
T,17,0.0%
VDS,1542112,1.0105%
S,700187,45.4%
V,255776,16.59%
P,173553,11.25%

```
npikkus=7
hoidla={}
nr=0
for rida in open("sonaliigid_koos.txt"):
  nr+=1
  if nr % 10000 == 0: print(nr)
  r=rida.strip().replace("Z", "")
  ngramid=[r[koht: koht+npikkus]
        for koht in range(len(r)-
npikkus+1)]
  for ngram in ngramid:
    if ngram in hoidla: hoidla[ngram]+=1
    else: hoidla[ngram]=1
```

Edasiarendusmõte

- Kui sarnane uurimine end õigustab, siis on mõtet luua ka graafiline liides, kus võimalik määrata kindlaid ja lisatunnuseid, järgnevusi ja kaasatulijaid ning alamlõike uurimiseks andmepuudena kokku ja lahku klõpsida.
- Katsetused erinevate keelekasutusvariantidega:
  - publitsistikas ja ilukirjanduses sageli esinevad keelekasutusmustrid ja nende tüüpilised kontekstid; sarnaste mustrite kasutuseelistused; keelevariandispetsiifilised mustrid ja kontekstid;
  - eesti õppijakeele mustrid ja kontekstid võrdluses emakeelekõneleja kirjutatud tekstides esinevate mustrite ja nende kontekstidega, mittenormatiivsete kontekstide automaatne markeerimine

Lugemist

- Allkivi, Kais 2016. C1-tasemega eesti keele õppijate ja emakeelekõnelejate kirjaliku keelekasutuse võrdlus verbialguliste tetragrammide näitel [Written language use of C1 learners of Estonian and native speakers in comparison: analysis of verb-initial fourgrams]. – Lähivõrdlusi. Lähivertailuja 26, 54–83. https://doi.org/10.5128/LV26.02
- Biber, Douglas 2009. A corpus-driven approach to formulaic language in English. – International Jpurnal of Corpus Linguistics 14 (3), 275–311. DOI: 10.1075/ijcl.14.3.08bib
- Chen, Yu-Hua & Baker, Paul 2010. Lexical bundles in L1 and L2 academic writing. – Language Learning & Technology 14 (2), 30–49. http://llt.msu.edu/vol14num2/chenbaker.pdf
- De Cock, Sylvie & Granger, Sylviane 2021. Stance in press releases versus business news: A lexical bundle approach. – Text and Talk 41 (5–6), 691–713.
- Eslon, Pille 2022. Leksika-grammatika piirimail: publitsistlike arvamuslugude keelekasutusmustrid [At the border of lexis and grammar: Language use patterns of journalistic opinion articles]. – Lähivõrdlusi. Lähivertailuja 32: 13–52.
- Eslon, Pille 2014a. Morfosüntaktilise ja leksikaalse varieerumise piiridest: ilukirjandus- ja õppijakeele kasutusmustrite võrdlus [Constraints on morphosyntactic and lexical variability]. – Eesti Rakenduslingvistika Ühingu aastaraamat 10, 55–71.

- Eslon, Pille 2014b. Adverbi sisaldavate struktuuride tekstifunktsioonidest eesti ilukirjandus- ja õppijakeeles [On the textual functions of adverbial structures in literary Estonian and in Estonian learner language]. – Lähivõrdlusi. Lähivertailuja 24, 15–46.
- Evert, Stefan 2005. The statistics of word cooccurrences, word pairs and collocations. PhD dissertation. Institut für maschinelle Sprachverarbeitung Universität Stuttgart.
- Jackendoff, Ray 2017 (2015). In defense of theory. – Cognitive Science 41 (Suppl. 2), 185–212. https://onlinelibrary.wiley.com/doi/10.1111/cogs.12324
- Juknevičienė, Rita 2013. Recurrent word sequences in written learner English. – Eds. I. Šeškauskienė & J. Grigaliūnienė. Anglistics in Lithuania. Cross-Linguistic and Cross-Cultural Aspects of Study. Cambridge Scholar Publishing, 178–197. https://www.academia.edu/11922891/Recurrent_Word_Sequences_in_Written_Learner_English
- Liiva, Kristjan 2022. Automaatne tekstianalüüs: Klastrileidja arendamise põhimõtted ja veebirakendus [The Clusterfinder Web Application and it's Development Principles]. Bakalaureusetöö. Tallinna Ülikool, Digitehnoloogiate instituut.
- Martinez, Ron & Schmitt, Norbert 2010. A phrasal expressions list. – Applied Linguistics 33 (3), 299–320.

- Matsak jt 2010 = Matsak, Erika & Eslon, Pille & Kippar, Jaagup 2010. Eesti keele sõnajärje vealeidja prototüübi arendamine [The development of the prototype for an automatic word order error detector for the Estonian language]. – Korpusuuring ja meetodid. Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 12. Tallinn: TLÜ EKKI, 59−100.
- Mel’čuk, Igor 1995. Semantics of two emotion verbs in Russian: *bojat’sja* '[to] be afraid’ and *nadejat’sja* ‘[to] hope’. Mel’chuk [Mel’čuk], Igor. Russkii iazyk v modeli ‘smysl □ tekst’. Moskva – Vena: Iazyki russkoi kul’tury, 81–133.
- Muischnek, Kadri 2006. Verbi ja noomeni püsiühendid eesti keeles [Fixed expressions consisting of verbs and nouns in Estonian]. Dissertationes Philologiae Estonicae Universitatis Tartuensis 17. Tartu: Tartu Ülikooli Kirjastus.
- Muischnek, Kadri & Kaalep, Heiki-Jaan 2010. The variability of multi-word verbal expressions in Estonian. – Language Resources and evaluation 44 (1–2), 115–135.
- Muischnek, Kadri & Sahkai, Heete 2010. Liitpredikaadid leksikoni-grammatika kontiinuumil: konstruktsioonide produktiivsusest verbiga *minema* moodustatud liitpredikaatide näitel [Complex Predicates on the Lexicon-Grammar Continuum]. – ESUKA/JEFUL 1 (2), 295–316.
- Ots, Sander 2012. Statistikapõhise tarkvara loomine morfoloogiliste kollokatsioonide eraldamiseks eesti keele tekstidest [Software for morphosyntactic cluster extraction from Estonian texts]. Bakalaureustöö. Tallinna Ülikool, informaatika instituut.

- Sahkai, Heete 2011. Eesti keele genitiivse agendifraasi süntaks [The syntax of the Estonian genitive agent phrase]. – Keel ja Kirjandus 1, 12−30.
- Sahkai, Heete 2008. Konstruktsioonipõhine keelemudel ja sõnaraamatumudel [Some lexicographic implications of a construction-based model of language]. – Eesti Rakenduslingvistika Ühingu aastaraamat 4, 177–186.
- Sahkai, Heete 2006. Konstruktsioonipõhise keelekirjelduse võimalustest adessiivse viisi- ja põhjusmääruse näitel [Konstruktionsbezogene Sprachbeschreibung am Beispiel der adessivischen Adverbialbestimmung der Art und Weise und des Grundes]. – Keel ja Kirjandus 10, 816–831.
- Schmitt, Norbert 2010. Researching vocabulary: A vocabulary research manual. – Eds. Christopher N. Candlin & David R. Hall. Research and practice in applied linguistics. Palgrave Macmillan.
- Selivan, Leo 2018. Lexical grammar. Activities for teaching chunks and exploring patterns (Cambridge Handbooks for Language Teachers). Cambridge University Press.
- Trainis, Jekaterina 2017. Diakroonilised nihked eesti ilukirjanduskeele kasutusmustrites 1890–1990 [Diachronic language changes in usage patterns of belletristic Estonian in 1890s–1990s]. – Mäetagused 69, 181−216. DOI: 10.7592/MT2017.69.trainis

- Trainis, Jekaterina & Allkivi, Kais 2014. Ilukirjanduskeelest uue pilguga [On belletristic language from a new perspective]. – Eesti Rakenduslingvistika Ühingu aastaraamat 10, 283–306. doi:10.5128/ERYa1018
- Õim, Asta & Õim, Katre 2019. Lähtekohti eesti fraseoloogia käsitlemiseks [On the starting-point of Estonian phraseology description]. Tallinn: EKSA.