TALLINNA ÜLIKOOL
Digitehnoloogiate instituut

TARTUNLP

Veamärgendusega korpus grammatikakontrollija arenduseks ja testimiseks

Kais Allkivi-Metsoja, Karina Kert, Silvia Maine – Tallinna Ülikool
Krista Liin – Tartu Ülikool

Taust

- TÜ ja TLÜ ühisprojekt „Eestikeelse teksti automaatkorrektuur“ (EKTB25)
- Vajadus:
  - standard automaatse veaparanduse hindamiseks
  - veastatistika ja täiendav treeningmaterjal grammatikakontrollija arenduseks
  - edasine automaatse veamärgendamise alus
- Lahendus:
  - olemasolevate tekstikogude põhjal koostatud ühtse veamärgendusega korpus, kus lausete kaupa tähistatud mitmesuguste keelevigade asukoht, liik ja parandus
  - sisaldab eesti keele kui teise keele õppijate ja eesti keelt emakeelena kõnelejate toimetamata kirjutisi
  - hõlmab üle 7000 lause – mahult võrreldav rahvusvahelise veaparandussüsteemide võistluse BEA-2019 arendus- ja testandmestikuga (Bryant et al., 2019)

Materjal

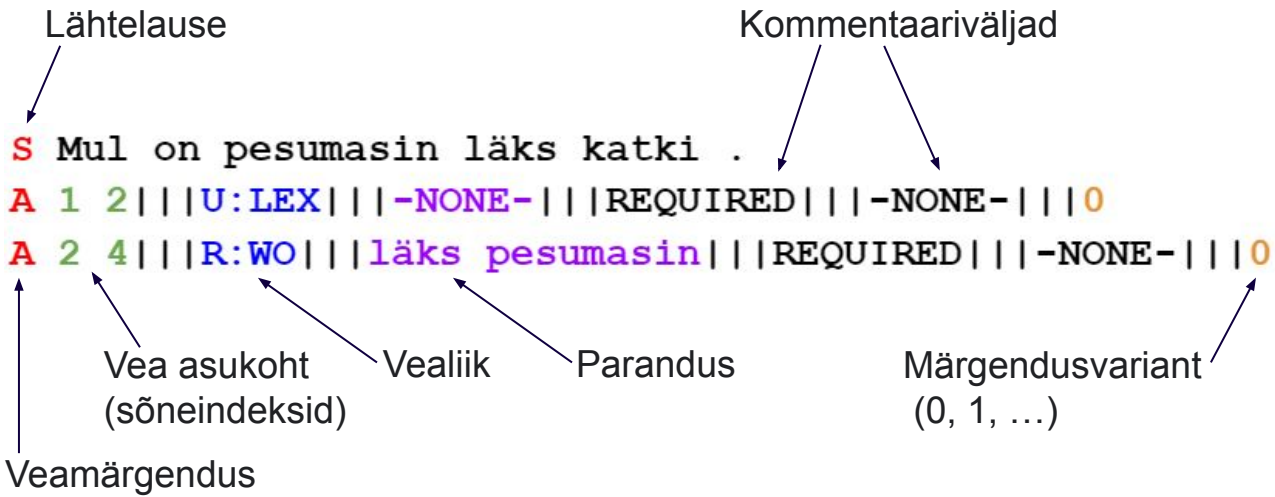
Eesti vahekeele korpus (EVKK)

- eesti keele tasemeeksamite kirjutised
- 263 teksti, 3790 lauset:
  - A2 – 940 lauset
  - B1 – 962 lauset
  - B2 – 1091 lauset
  - C1 – 796 lauset
- eelnevalt märgendatud CoNLL-U formaadis, tähistatud õigekirja-, grammatika- ja sõnastusvead
- üks parandusversioon lisatud kahe märgendaja koostöös

Eesti keele õppija korpus EMMA

- eesti emakeelega keskkooliõpilaste eksamikirjutised
- 81 teksti, 3546 lauset
- eelnevalt märgendatud õigekirja-, kirjavahemärgi-, stiili- ja faktivead
- üks märgendusversioon (parandusteta) lisatud kahe märgendaja koostöös

Märgendusformaat (M2)



Formaadi võtsid kasutusele Dahlmeier ja Ng (2012), kes pakkusid välja veaparanduse hindamise algoritmi MaxMatch (M^2).

Vealiigitus

Asendus (R – replacement)

- õigekirjaviga (R:SPELL)
- algustäheviga (R:CASE)
- kokku-lahkukirjutamise viga (R:WS)
- käändsõna vormivaliku viga (R:NOM:FORM)
- tegusõna vormivaliku viga (R:VERB:FORM)
- sõnavalikuviga (R:LEX)
- sõnajärjeviga (R:WO)
- kirjavahemärgi valiku viga (R:PUNCT)

Puudumine (M – missing)

- puuduv sõna (M:LEX)
- puuduv kirjavahemärk (M:PUNCT)

Liiasus (U – unnecessary)

- liigne sõna (U:LEX)
- liigne kirjavahemärk (U:PUNCT)

Eesti keelele kohandatud ERRANT-i veaklassifikatsioon (Bryant et al., 2017).

Algse lause päästmine

Näide 1. Minimaalse paranduste arvuga saavutatav grammatiline lause stiilivigadele tähelepanu pööramata

S Mulle meeldib see **auto** , aga ma ostsin uut **autot** .
Mulle meeldib see auto , aga ma ostsin uue auto .
A 8 9|||R:NOM:FORM|||uue|||REQUIRED|||-NONE-|||0
A 9 10|||R:NOM:FORM|||auto|||REQUIRED|||-NONE-|||0

Leksikaalsed valikud

Näide 2. Kontekstis sobimatu sõna kasutamine

S Ta oli ebanormaalne õpilane .
Ta oli ebatavaline||ebaharilik õpilane .
A 2 3|||R:LEX|||ebatavaline||ebaharilik|||REQUIRED|||-NONE-|||0

Liitmärgendid

Näide 3. Samas kohas on korraga mitu viga, mida on parandatud liitmärgendiga

S See on pealinn Islandil .
See on Islandi pealinn .
A 2 4|||R:WO:NOM:FORM|||Islandi pealinn|||REQUIRED|||-NONE-|||0

12 põhimärgendi kõrval võtsime kasutusele 18 liitmärgendit.

Mitu parandusvõimalust

Näide 4. Lähtelauses on vale ajavorm, mida saab parandada kahel viisil

S Olime õppinud koolis 12 aastat .

Oleme õppinud koolis koos 12 aastat.
A 0 1|||R:VERB:FORM|||Oleme|||REQUIRED|||-NONE-|||0
A 3 3|||M:LEX|||koos|||REQUIRED|||-NONE-|||0

Õppisime koolis koos 12 aastat .
A 0 2|||R:VERB:FORM|||Õppisime|||REQUIRED|||-NONE-|||1
A 3 3|||M:LEX|||koos|||REQUIRED|||-NONE-|||1

Emakeeleõppija kirjanditekstid

- Esialgsed märgendid olid aluseks, ent võis lisada juurde.
- Mitte kõik algsed vead ei saanud märgendatud:
  - Taandread, poolitusvead, nt “*tavalise-lt*”
  - Faktivead, stiilivead, nt *“roosilise eluga”*
  - Võimalikud ära parandatud vead, nt suurtähestusviga lause algul
- Muud probleemid
  - Võimalikud sisestusvead, nt “*inmene*”, “*tavakodaniku/el*”
  - Lausestusvead, nt “*A. H. Tammsaare*”
  - Õigekirjareeglid ja -soovitused on viimase 25 aasta jooksul muutunud.

Emakeeleõppija veakohad

- Sõnavalik
- Paronüümid
- Sõnakordused
- Seotud vead
- Otsekõne
- Kirjavahemärk-sõna vahetus

Veastatistika

Märgendaja 0 järgi arvestades

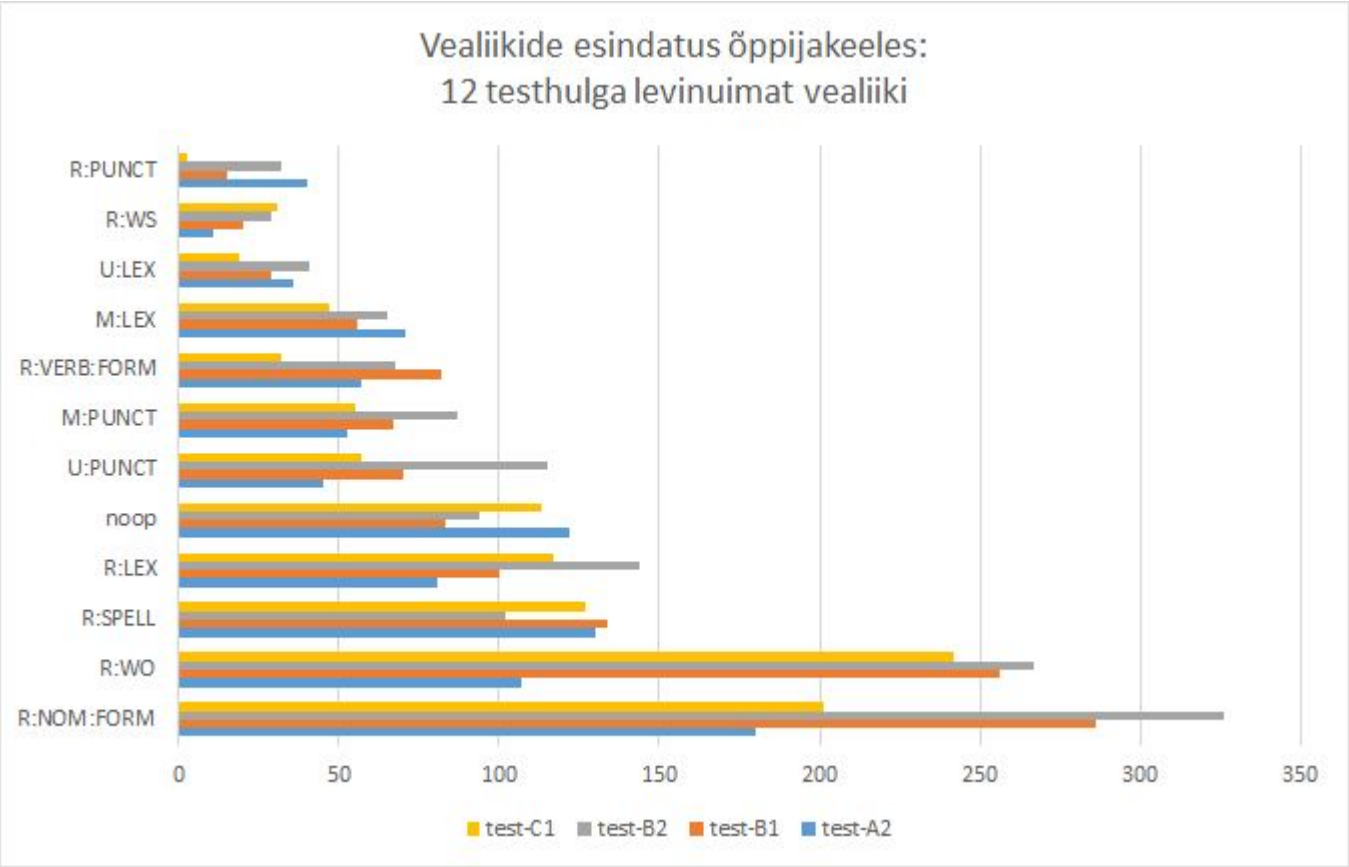
	EVKK dev	EVKK test	EMMA test
Lauseid kokku	257	2029	1449
Korrektseid lauseid	22%	20%	69%
Vigadega lauseid	78%	80%	31%
Vigu kokku	537	4392	450

Levinumad vealiigid testkorpuses

EVKK A2		EVKK B1		EVKK B2		EVKK C1		EVKK kokku		EMMA	
R:NOM:FORM	20%	R:NOM:FORM	24%	R:NOM:FORM	24%	R:WO	25%	R:NOM:FORM	23%	noop	222%
R:SPELL	15%	R:WO	21%	R:WO	20%	R:NOM:FORM	21%	R:WO	20%	M:PUNCT	27%
noop	14%	R:SPELL	11%	R:LEX	11%	R:SPELL	13%	R:SPELL	11%	R:LEX	24%
R:WO	12%	R:LEX	8%	U:PUNCT	9%	R:LEX	12%	R:LEX	10%	U:PUNCT	19%
R:LEX	9%	noop	7%	R:SPELL	8%	noop	12%	noop	9%	R:WS	14%

Vealiik *noop* näitab korrektsete lausete suhet vigade koguhulka
Protsendid on arvutatud seda välja jättes.

Kõik vealiigid olid kasutusel, mõni seni vaid ühel korral (nt R:WS:NOM:FORM:CASE).


Vealiikide esindatus õppijakeeles:
12 testhulga levinuimat vealiiki
R:PUNCT
R:WS
U:LEX
M:LEX
R:VERB:FORM
M:PUNCT
U:PUNCT
noop
R:LEX
R:SPELL
R:WO
R:NOM:FORM
0
50
100
150
200
250
300
350
test-C1
test-B2
test-B1
test-A2

M^2 hindamisskript

- MaxMatch (M^2) hindaja
  - Eeldab M2 märgendusformaati.
  - Iga lause puhul:
    - Leiab algse lause ja parandatud lause põhjal vähimad võimalikud parandused.
    - Võrdleb neid iga algse kuldstandardi märgendajaga eraldi.
    - Valib parima F-skooriga märgendaja.
  - Väljastab kumulatiivse täpsuse ja saagise pakutud paranduste kohta.
- Korpusest tulenevad muudatused:
  - WO veamärgend
    - pikema veaskoobi lubamine
    - kattuvate väiksemate vigadega arvestamine
  - Tulemused vealiikide kaupa

M^2 hindamisskript: teisi hõlmav vealiik WO

S Mul ei **olnud** kassi mitte kunagi .

A 2 3|||R:VERB:FORM|||**ole olnud**|||REQUIRED|||-NONE-|||0

A 2 6|||R:WO|||**ole** mitte kunagi kassi **olnud**|||REQUIRED|||-NONE-|||0

Kaks lisamärgendajat:

- algne märgendaja 0, kellel on vahemikus 2–6 **kaks kattuvat parandust.**
- pseudomärgendaja 1, kellel on vahemiku 2–6 jaoks vaid üks suur **sõnajärjeparandus**.
- pseudomärgendaja 2, kellel on vahemiku 2–6 jaoks alles vaid **pisemad parandused**.

M^2 hindamisskript: tulemused vealiikide kaupa

Esimene samm ERRANTi hindamisskripti kohandamise suunas.

- Iga vealiigi kohta:
  - Koguarv – muutub vastavalt valitud märgendajale
  - Saagis
    - veaparandusel
    - veatuvastusel (täpsetes piirides)
    - veatuvastusel (osalistes piirides)

- Kaks erinevat statistikat WO vealiiki arvestades

M^2 hindamisskript: tulemused vealiikide kaupa

Saagis vealiigiti:

vealiik	kokku	parandatud	tuvastatud	kattuv_parandus
R:VERB:FORM	56	0.34	0.34	0.93
R:WO	240	0.34	0.36	0.95
R:LEX	143	0.27	0.27	0.87

Saagis vealiigiti, WO sees olevaid vigu arvestades:

vealiik	kokku	parandatud	tuvastatud	kattuv_parandus
R:VERB:FORM	69	0.46	0.46	0.94
R:WO	240	0.34	0.36	0.95
R:LEX	158	0.34	0.34	0.88

Täname

Kaisa Norak ja Pille Eslon (TLÜ) – märgendusskeemi täpsustamine, probleemide arutelu

Kirjandus

Bryant, C., Felice, M., Andersen, O. E., & Briscoe, T. (2019). The BEA-2019 Shared Task on Grammatical Error Correction. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (lk 52–75). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W19-4406

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada* (lk 793–805). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1074

Dahlmeier, D., & Ng, H. T. (2012). Better Evaluation for Grammatical Error Correction. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, Canada* (lk 568–572). Association for Computational Linguistics. https://aclanthology.org/N12-1067.pdf