TALLINNA ÜLIKOOL
Digitehnoloogiate instituut

TARTU NLP

Eesti keele õigekirja- ja grammatikakontroll: mudelite võrdlus ja kombineerimine

Agnes Luhtaru, Krista Liin, Mark Fišel – Tartu Ülikool
Kais Allkivi-Metsoja, Jaagup Kippar – Tallinna Ülikool

Projektist

- TÜ ja TLÜ ühisprojekt EKTB25 „Eestikeelse teksti automaatkorrektuur“ (2021–2023)
- Eesmärk: arendada edasi eesti keele õigekirja- ja grammatikakontrolli vahendeid ja need omavahel ühendada
- Põhimeetodid: statistikapõhine õigekirjakontroll ja neuromasintõlkele tuginev grammatikakontroll
- Korrektuurivahendite komplekt koodivaramus:
  - kaks grammatika- ja kolm õigekirjakontrolli mudelit
  - mudeleid saab kasutada ühekaupa või jadana
  - https://koodivaramu.eesti.ee/tartunlp/corrector

Statistiline õigekirjakontroll

- Keelest sõltumatu ja vähenõudlik kontekstitundliku veaparanduse lahendus
- Vajab treeningandmetena vaid korrektse keelekasutuse näiteid
- Hindab võimalike paranduste tõenäosust n-grammide ehk sõnajärjendite sageduse alusel
- Kolm katsetatud algoritmi:
  - **Norvig** – genereerib parandusi tähtede lisamise, kustutamise, asendamise ja nihutamise teel; algselt tugines unigrammidele, meie kasutasime paranduste reastamiseks bigrammide andmeid
  - **Symspell** – taandab võimalikud muudatused tähekustutusteks; lähtub bigrammidest ainult juhul, kui unigrammide loendist väga sarnast vastet ei leia
  - **Jamspell** – Symspelli optimeeritud mälukasutusega edasiarendus, arvestab paranduste hindamisel bi- ja trigrammidega

Spellerite võrdlus

- Eeldus: statistiline õigekirjakontroll võimaldab keeleõppijate vigu täpsemini tuvastada ja parandada kui sõnastiku- ja reeglipõhine lähenemine
- Võrreldavad spellerid: Vabamorf, MS Wordi ja Google Docsi korrektor
- Treeningmaterjal: ühendkorpus 2019 -> koondkorpuse valim
- Testmaterjal: EVKK eesti keele A2–C1-taseme eksamite loovkirjutised
  - 1054 lauset ja 9186 sõna (u 2000–3000 sõna keeleoskustaseme kohta)
  - 309 õigekirjaviga, mis 46 juhul kattusid muu keeleveaga (algustäht, vormi/sõnavalik)
  - CoNLL-U formaadis veamärgendus, kus tähistatud eri liiki vead koos parandusega
- Hindamine:
  - õigekirjavigade tuvastamine ja parandamine
  - arvesse võetud esimene parandusvariant, täielikud ja osalised parandused (ainult õigekiri)

Spellerite võrdlus: veatuvastus (%)

Speller	F0,5	Täpsus	Saagis
Jamspell	**83,9**	**89,6**	**67,0**
Norvig	78,9	84,3	62,8
Symspell	69,1	86,2	38,5
Vabamorf	**84,3**	**89,2**	69,3
MS Word	83,4	87,8	**69,6**
Google	76,7	78,8	**69,6**

Spellerite võrdlus: veaparandus (%)

Speller	F0,5	Täpsus	Saagis
Jamspell	**64,1**	**68,4**	**51,1**
Norvig	54,1	57,8	43,0
Symspell	31,4	39,1	17,5
Google	**67,5**	**69,2**	**61,2**
MS Word	51,2	53,9	42,7
Vabamorf	42,6	45,0	35,0

Spellerite võrdlus: järeldused

- Jamspell ja bigrammipõhine Norvigi speller edestasid veatuvastuses Google'i korrektorit ning veaparanduses Vabamorfi ja MS Wordi korrektorit
- Jamspell saavutas veatuvastuses ligilähedase tulemuse Vabamorfiga ja veaparanduses kõige sarnasema tulemuse Google'iga
- Google tegi enim tarbetuid parandusi (21,2%), Jamspell ja Vabamorf tegid neid kõige vähem (vastavalt 10,4% ja 10,8%)
- Jamspell ja Norvigi speller pakkusid kontekstitundlikke parandusi (nt **tõdida ~ tõdeda*, mitte *tüdida*; **ludeda ~ lugeda*, mitte *kudeda*); Jamspell tuli kõige paremini toime homonüümsete vigadega (nt **vaga ~ väga*, **kuued ~ kuud*)

Jamspelli mudelite võrdlus: treeningandmed

Treeningandmed	Lauseid	Sõnu
Veebikorpus 2019	40,9 mln	512,6 mln
Koondkorpus + Vikipeedia korpused + DOAJ	16,9 mln	230,1 mln
Koondkorpus	13,2 mln	180,9 mln
Koondkorpuse valim	6 mln	82,4 mln
Veebikorpuse valim	6 mln	75,2 mln
Koondkorpuse + veebikorpuse 10:1 valim	6,6 mln	89,9 mln
Koondkorpuse + veebikorpuse 1:1 valim	6 mln	78,9 mln
Koondkorpuse + Vikipeedia + DOAJ valim	4,2 mln	55,7 mln

Jamspelli mudelite võrdlus: veatuvastus (%)

Treeningandmed	**F0,5**	**Täpsus**	**Saagis**
Koondkorpuse valim	**83,9**	89,6	**67,0**
Koondkorpus + veebikorpus 10:1	82,7	91,2	60,2
Veebikorpus 2019	81,9	**94,3**	53,7
Koondkorpus + Vikipeedia + DOAJ valim	80,4	87,7	60,2
Koondkorpus + veebikorpus 1:1	79,9	89,6	55,7

Jamspelli mudelite võrdlus: veaparandus (%)

Treeningandmed	**F0,5**	**Täpsus**	**Saagis**
Veebikorpus 2019	**64,7**	**74,4**	42,4
Koondkorpuse valim	64,1	68,4	**51,1**
Koondkorpus + Vikipeedia + DOAJ valim	63,5	69,3	47,6
Koondkorpus + veebikorpus 10:1	63,1	69,6	46,0
Koondkorpus + veebikorpus 1:1	63,1	70,8	44,0

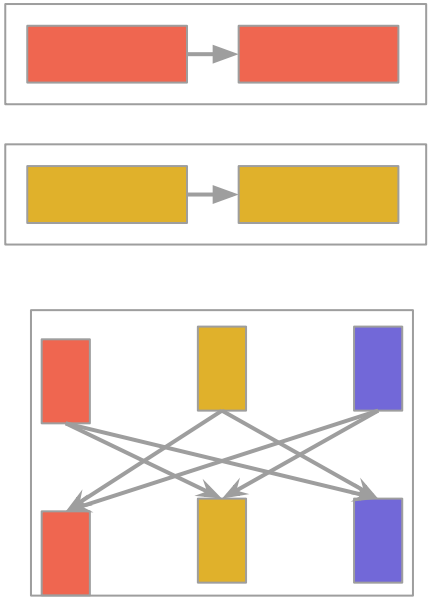
Jamspelli mudelite võrdlus: järeldused

- Veatuvastuses ja -paranduses oli täpseim veebikorpusel treenitud mudel, mis tegi vähim tarbetuid parandusi, kuid leidis ka kõige vähem vigu
- Suurima saagise saavutas koondkorpuse valimil treenitud mudel, mille täpsus oli tarbetute paranduste arvelt tagasihoidlikum
- Standardsema treeningkorpuse suurendamine ei paranda tulemusi, mitmekesisema veebikorpuse puhul annab suurem maht eelise
- Kompromissi – vahepealset täpsust ja saagist – pakub mudel, mis treenitud suuremal määral koondkorpuse ja vähemal määral veebikorpuse materjaliga
- Mudeli valik sõltub vajadusest

Grammatiliste vigade parandamine

Kaks erinevat metoodikat, mis mõlemad põhinevad neuromasintõlkel

1. Grammatikaparandus **kui ühesuunaline masintõlge** → õpetame mudeli tõlkima vigasest tekstist korrektsesse teksti
2. Grammatikaparandus **kui mitmekeelse masintõlke lisaoskus** → erinevate keelte vahel tõlkima õpetatud mudel parandab ka vigu



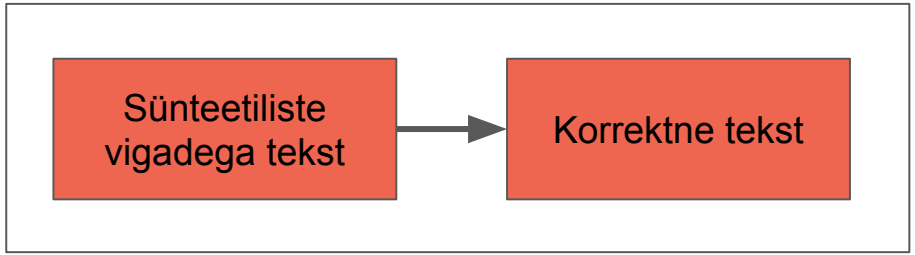
Peamine raskuskoht on andmete vähesus

- Meetodite valikus fookus:
  Kuidas kasutada olemasolevaid ressursse, näiteks märgendamata teksti ja tõlkenäiteid?

Keel	inglise	eesti
Kasutatavate veaparandusnäidete arv (lausetes)	> 1 000 000	~ 7 000

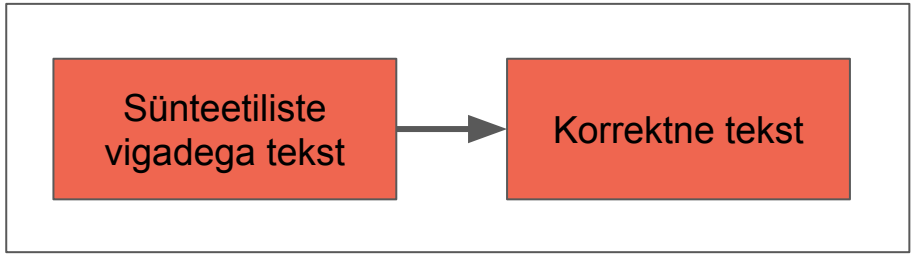
Meetod 1: Sünteetiliste vigadega eeltreenimine

Eeltreenimise samm, ainult sünteetilised andmed

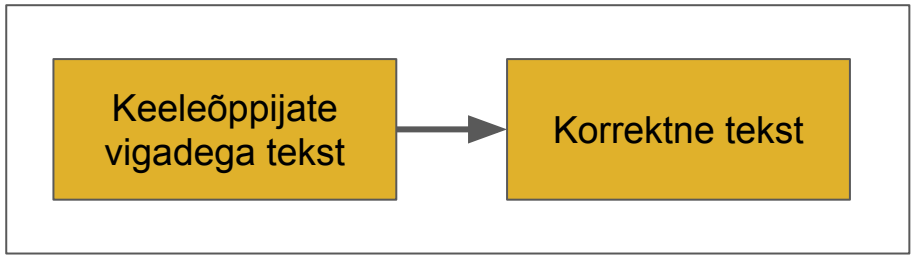


Meetod 1: Sünteetiliste vigadega eeltreenimine

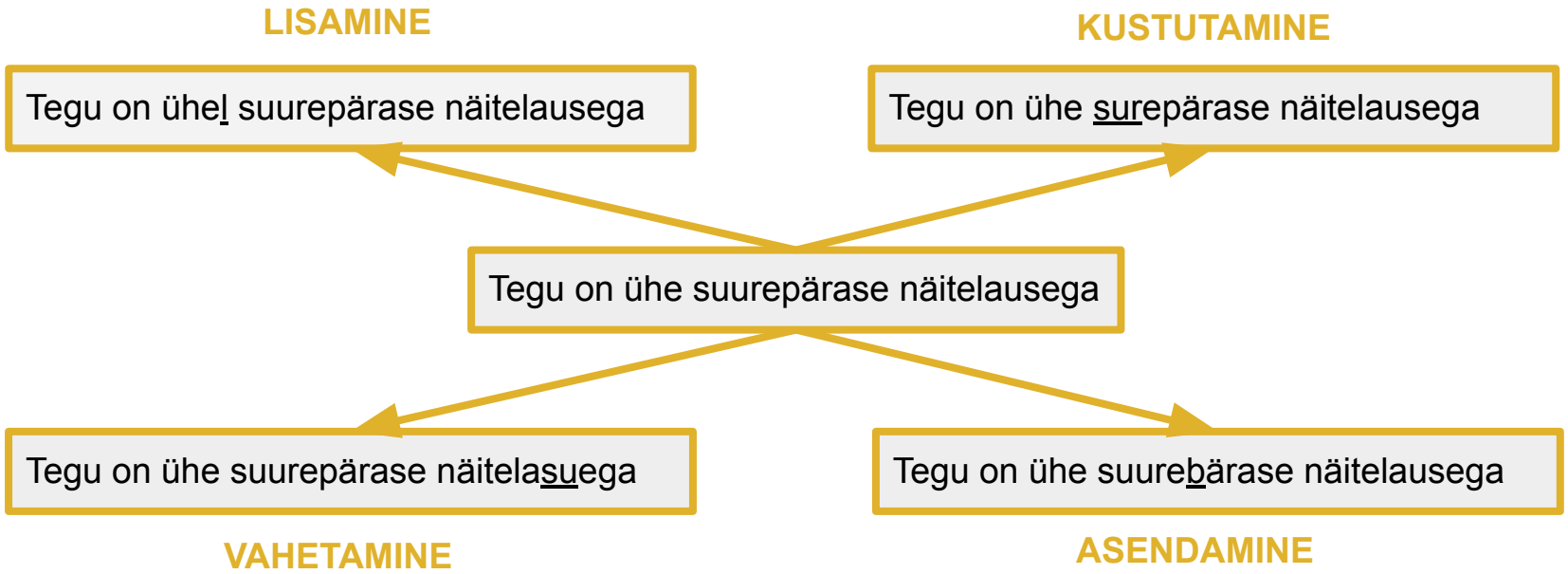
Eeltreenimise samm, ainult sünteetilised andmed



Häälestamise samm, näited inimeste kirjutatud tekstidest



Sünteetilised vead



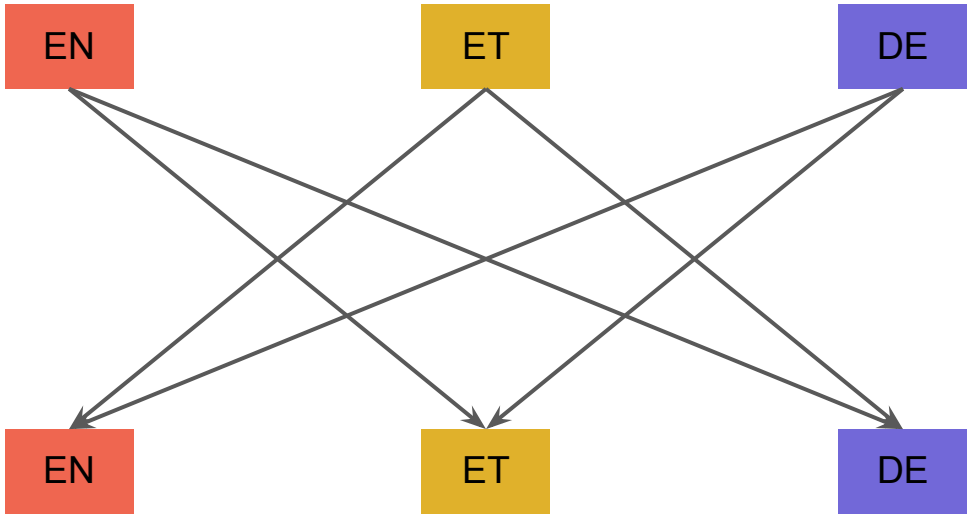
Sõnade asendamiseks tagurpidi spelleri meetod

- Asendame spelleri (Aspell) pakutud variantidest suvalise sõnaga

Sõna	Pakutavad asendused
kunst	kunts, kunste, kunsti, kust
kuivatatud	kuivatanud, kuivatus, kuivatama
sõlmida	sõlmuda, sõlmid, salmida
säästva	säästvat, säästa, päästva
eksperdi	eksperdid, eksperdil, eksperdis

Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data (Grundkiewicz et al., BEA 2019)

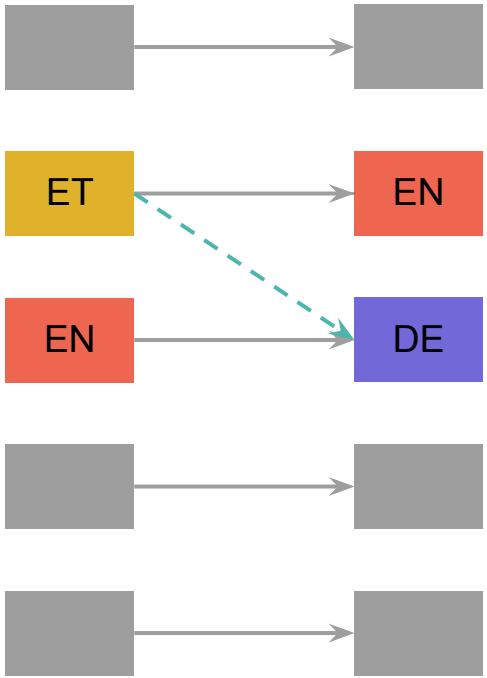
Meetod 2: Mitmekeelne masintõlge



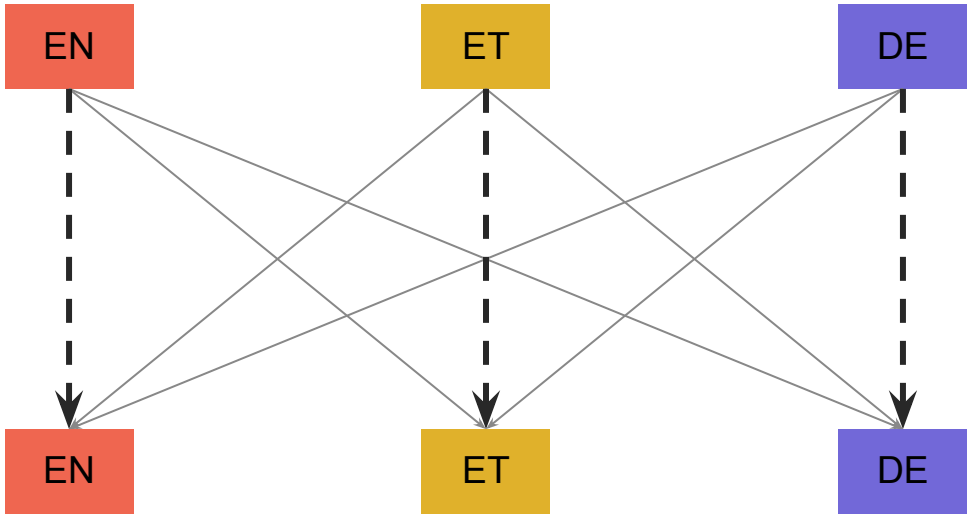
Zero-shot tõlge

Tõlge keelepaaride vahel,
mida mudel pole näinud.

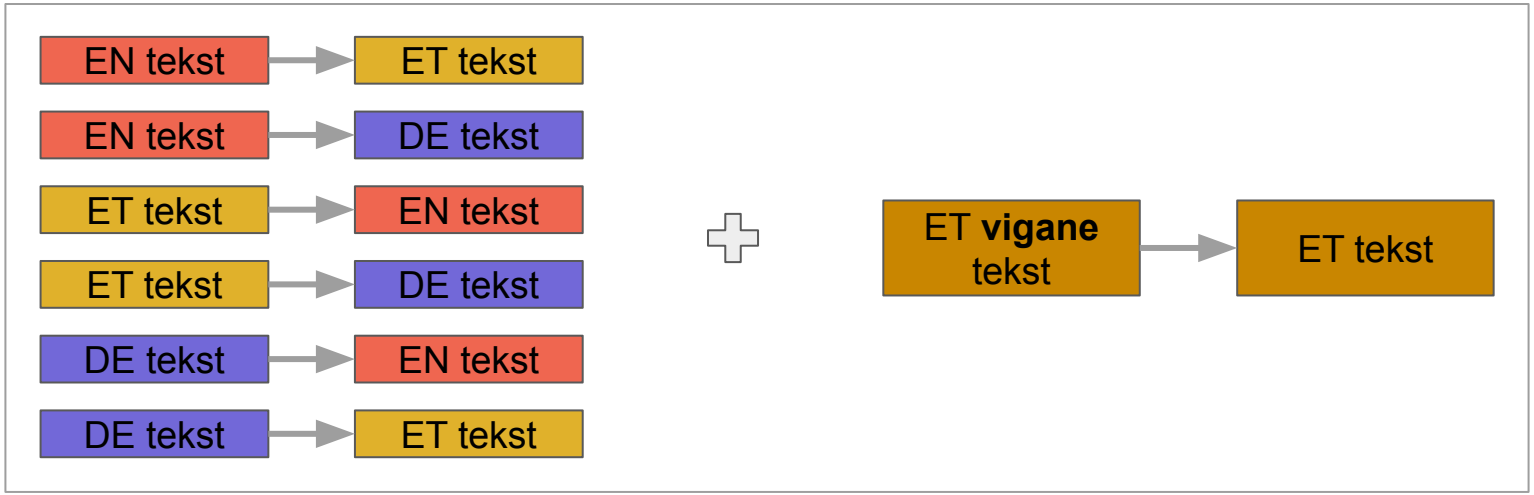
Treenimisel: ET → EN & EN → DE
Kasutamisel: ET ⇢ DE



Meetod 2: Ühekeelne *zero-shot* tõlge parandab vigu



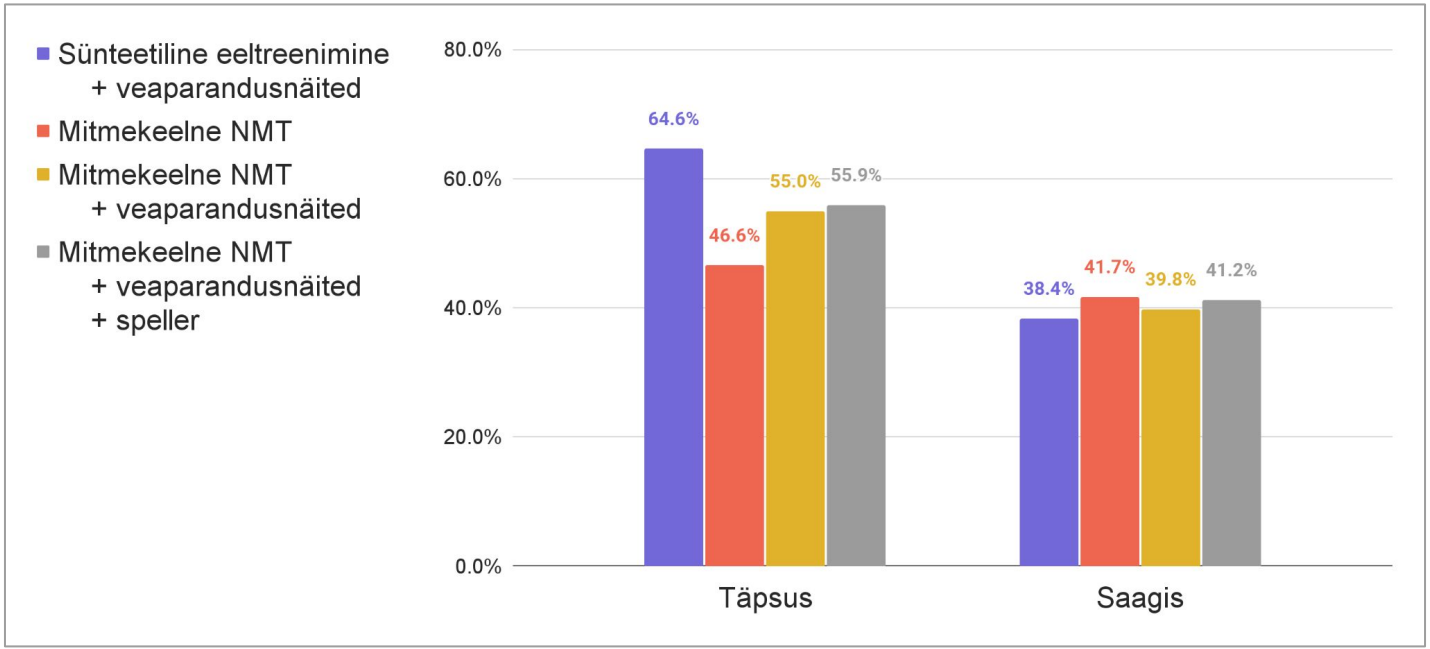
Tõlkemudelit saab edasi treenida veaparandusnäidetega



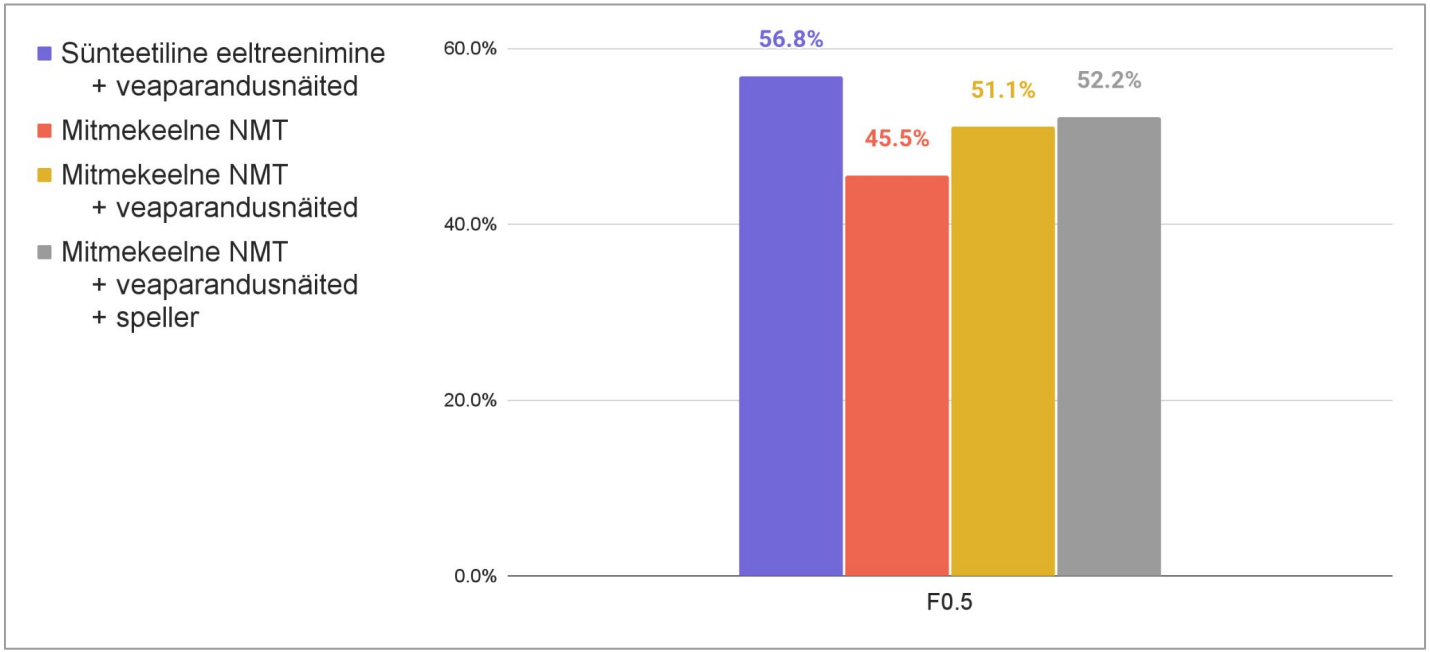
Mudelite hindamine

- Üksi ja kombineeritult (speller + grammatikakontroll)
- Testmaterjal: 2029 lauset eesti keele tasemeeksamite kirjutistest
  - A2 – 495, B1 – 504, B2 – 534, C1 – 495 lauset
  - M2-formaadis veamärgendus, lause kohta kuni kolm parandusvarianti
- Rakendus M^2 Scorer
  - arvutab veaparanduse täpsuse, saagise ja F-skoori, lähtudes iga lause puhul parandusvariandist, millega korrektori väljund on kõige sarnasem
  - kohandatud eesti keelele: sõnajärjevigade skoop võib kattuda teiste vigadega
  - täiendatud versioon arvutab saagise vealiikide kaupa

Parimad kombinatsioonid: täpsus ja saagis



Parimad kombinatsioonid: F0,5-skoor



Veaparanduse saagis vealiigiti

Koondkorpusel treenitud speller + mitmekeelne NMT

1. puuduv kirjavahemärk – 80%
2. õigekiri – 67%
3. algustähe valik – 60%
4. tegusõna vormivalik – 56%
5. käändsõna vormivalik – 54%
6. liigne sõna – 51%
7. sõnajärg – 47%
8. kokku-lahkukirjutamine – 40%
9. sõnavalik – 39%
10. puuduv sõna – 35%
11. liigne kirjavahemärk – 32%
12. kirjavahemärgi valik – 5%

Sünteesvigadel eeltreenitud grammatikakontrollija

1. õigekiri – 76%
2. puuduv kirjavahemärk – 70%
3. tegusõna vormivalik – 51%
4. käändsõna vormivalik – 49%
5. liigne sõna – 49%
6. sõnajärg – 47%
7. algustähe valik – 42%
8. puuduv sõna – 37%
9. sõnavalik – 30%
10. liigne kirjavahemärk – 30%
11. kokku-lahkukirjutamine – 26%
12. kirjavahemärgi valik – 3%

Näited: grammatikakontroll

1. Natuke **kõike need probleemi aitab** lahendada kiire tehnoloogia areng.

   Sünteesvigadel eeltreenitud grammatikakontrollija:
   Natuke **aitab kõiki neid probleeme** lahendada kiire tehnoloogia areng.

   Mitmekeelne NMT:
   Natukene aitab [...] neid probleeme lahendada kiire tehnoloogia areng.

2. Ma tahan külastada **tartu**.

   Sünteesvigadel eeltreenitud grammatikakontrollija:
   Ma tahan külastada **tartut**.

   Mitmekeelne NMT:
   Ma tahan **Tartut külastada**.

Näited: speller + grammatikakontroll

1. Koera kohta ma arvan, et ma võin osta **poodist** või küsida sõprade **juurest**, sest nad võivad aidata valida **soobiva** koera.

   Speller:
   Koera kohta ma arvan, et ma võin osta **poodist** või küsida sõprade **juurest**, sest nad võivad aidata valida **sobiva** koera.

   Mitmekeelne NMT:
   Koera kohta ma arvan, et ma võin osta **poest** või küsida sõprade **käest**, sest nad võivad aidata valida **sooviva** koera.

   Speller + mitmekeelne NMT:
   Koera kohta ma arvan, et ma võin osta **poest** või küsida sõprade **käest**, sest nad võivad aidata valida **sobiva** koera.

2. Ma läksin **musemius** ja **ekskursioni**.

   Speller:
   Ma läksin **muuseumi** ja **ekskursiooni**.

   Sünteesvigadel eeltreenitud grammatikakontrollija:
   Ma läksin **muuseumiks** ja **ekskursioni**.

   Speller + sünteesvigadel eeltreenitud grammatikakontrollija:
   Ma läksin **muuseumi** ja **ekskursiooni**.

Perspektiivid

- Testimine emakeelekõnelejate tekstidega (märgendamisel)
- Osaliste, vale- ja tarbetute paranduste analüüs

- Grammatikamudelite kombineerimine teistel viisidel
- ChatGPT, GPT-4 jt hindamine veaparanduseks
- Katsetused sünteetiliste vigade loomisel (nt veastatistikaga arvestamine, edasi-tagasi masintõlge)
- Reeglite lisamine parandustäpsuse suurendamiseks
- Õigekirjakontrolli täiendamine nimetuvastuse ja vealoendi abil