

E2 õppijamaterjali süntaktilise parsimise hindamiskorpuse koostamine ja parsimiskvaliteedi hindamise metoodika väljatöötamine

Tellija: Eesti Keele Instituut

Hanke nimetus: E2 õppijamaterjali süntaktilise parsimise hindamiskorpuse koostamine ja parsimiskvaliteedi hindamise metoodika väljatöötamine

Hanke ja selle eesmärgi lühikirjeldus:

Haridus- ja Teadusministeeriumi ning Euroopa Liidu kaasrahastatud programmi “Eesti keele õpe ja keeleõppe arendamine” raames tehakse ettepanek esitada pakkumine hankele “E2 õppijamaterjali süntaktilise parsimise hindamiskorpuse koostamine ja parsimiskvaliteedi hindamise metoodika väljatöötamine”. Hankes annab Tellija Täitja jaoks kasutada E2 õppijate (eesti keel teise keelena õppijate) veamärgendatud andmestiku, mille alusel Täitja läbib hankes ettenähtud tegevused ja tulemusena saavutab andmestiku ja metoodika, mida oleks võimalik kasutada grammatikaparserite soorituse hindamiseks E2 materjalide parsimise puhul. Metoodiliselt töötatakse välja nii hindamis- kui ka benchmarkimisalused, kuid rakendatavate benchmarkide väljatöötamist ei eeldata.

1. Hanke sisu

Hanke teostamise sisendina annab Tellija enda poolelt Täitjale kasutada alusandmestiku, mille näol on tegemist alaosa EKI error-annotated Estonian L2 learner corpus (version 2)¹ õppijakeele põhiseisveamärgenduskorpusest. See on salvestatud vrt-formaadis ja sisaldab süntaktilist infot (parser: estonian-edt-ud-2.15-241121) ja veamärgendusinfot (ERRANT-M2² veamärgendusformaadil põhinevat).

Hange teostatakse kahes etapis, millest esimene sisaldab Tellija pakutavast Alusandmestikust lähtuvalt teisenduskoodi kirjutamist (kood on Tellijale üleantavate tööde osa) ja selle rakendamisel esmalt Süntaktilise märgenduseta võrdlusandmestiku (vordlusandmestik_synt_ilma) genereerimist ja seejärel vordlusandmestik_synt_ilma põhjal Süntaktilise automaadmärgendusega võrdlusandmestiku tekitamist (vordlusandmestik_synt_auto). Andmestike vormi ja nendega teostatavate tegevuste

¹ <https://doi.org/10.15155/5hwt-3n69>

² <https://github.com/tlu-dt-nlp/EstGEC-L2-Corpus>

esitatavad täpsemad eeldused on esitatud peatüki "Protsessielemendid (andmestikud, kood, metoodika) ja protsesside kirjeldus" all.

Hanke teostamise teises etapis viiakse läbi võrdlusandmestiku_synt_auto grammatilise märgenduse ühestamine. Sellega töötatakse välja Süntaktilise käsitsimärgendusega võrdlusandmestik (vordlusandmestik_synt_yhest). Selle andmestiku näol ongi tegemist hindamisandmestikuga, mille alusel peab olema võimalik erinevate grammatikaparserite sooritust E2 õppijakeelematerjali parsimisel olema võimalik hinnata.

Selleks, et hindamisandmestikku vordlusandmestik_synt_yhest oleks võimalik eesmärgipäraselt tõhusalt rakendada, töötab Täitja hanke teostamise teises etapis välja ka metoodilised alused, mille alusel korpus parsimise hindamiseks ja benchmarkimiseks rakendada. Täpsustused metoodika osas on kirjeldatud peatüki Metoodika all.

2. Protsessielemendid (andmestikud, kood, metoodika) ja protsesside kirjeldus

Tööde teostumise I järgus

- Alusandmestik

- Formaat: vrt-formaadis. Sisaldab nii grammatilist kui ka veamärgenduslikku infot (estonian-edt-ud-2.15-241121, ERRANT-M2). Alusandmestiku maht on ligikaudu **2000 lauset**.
- Moodustumine: Võimaldab kasutamiseks Tellija.
- Eesmärk: Andmestikus on olemas veamärgenduslik info, mis on alus selleks, et oleks võimalik selle alusel välja töötada nõutav andmestik ja metoodika.
- Millele tähelepanu juhtida: veamärgendamisel kasutatud natuke täiendatud kategooriaid võrreldes algse ERRANT-M2 skeemiga. Tellija võimaldab koos korpusega ka selle kohta vajaliku info.

- Kood

- Formaat: Täitja otsustada. Peab lähtuma headest koodikirjutamise standarditest ja migreeritav Tellija koodireposse (GitHub).
- Moodustumine: Luuakse täitja poolt. Tellija on valmis andma eeskujuna kasutada ka enda kirjutatud sarnast eesmärki täitva, kuid puudujääkidega koodi.
- Eesmärk: Koodi alusel viia läbi teisendus Alusandmestikust Süntaktilise märgenduseta võrdlusandmestikku.
- Millele tähelepanu juhtida: näiteks võib keerulisust tekitada kattuvate *span*idega veaparanduskohtade käsitlemine. Töö teostamise jooksul on

Tellija valmis küsimusi tekitavad kohad jooksvalt kahepoolsest läbi arutama.

Tähelepanu tasub pöörata ka sellele, kas võrdlusandmestiku loomisel kuidagi ära markeerida need kohad, kus võrreldavad lausekujud teineteisega kattuvad. Kuigi genereeritud peavad saama kõik nõutud lausekujud ja järgnevate protsesside käigus peavad küll märgendused ja ühestamised külge saama kõik need kujud, võib samasuste kätteviiitamine välja tuua, kus on võimalik vältida topelttööd ja mingit osa tööprotsessist optimeerida.

- **Süntaktilise märgendusega võrdlusandmestik** (vordlusandmestik_ilma_synt)
 - **Formaat:** Alusandmestiku (vrt) põhjal tuletatakse lausete kohta paralleelvormid, mis väljendavad järgmist:
 - Lause algsel kujul nii nagu see oli keeleõppija kirjutatud (koos vigadega).
 - Lause korrigeeritud kujul, mis on tuletatud veaparanduste alusel kõiki veaparandusi arvessevõtvalt.
 - Lause osaliselt korrigeeritud kujul, mis on tuletatud veaparandusi ja nende kombinatsioone (max 3 suurusega) valikuliselt välja jättes. See info on oluline hilisemaks faasiks, kui on vaja võimaldada tuletada, kas parsimisvea puhul on tegemist parserist tuleneva parsimisveaga või keeleveast tuleneva parsimisveaga ning milline on keelevigade või nende kombinatsioonide mõju parsimistulemusele.
 - **Moodustumine:** Genereeritakse Tellija pakutud Alusandmestikust Täitja loodud koodi alusel.
 - **Eesmärk:** Alusandmestikust genereerida kõrvutatavad lausekujud, mis on salvestatud süstemaatiliselt ja olulist infot väljendavalt (nagu lausekuju moodustumise alused). Kasutatakse sisendiks järgmise teisendusjärgu jaoks, kus lausekujud saavad automaatparseriga külge süntaktilise märgenduse.
 - **Millele tähelepanu juhtida:** Soovitav on andmestiku salvestamisel tekitada nii lausetele tekstifailid kui ka metaandmefail (näiteks tabel), millesse kogu oluline info lausekujude moodustumise jms kohta jäädvustada.
- **Süntaktilise automaatmärgendusega võrdlusandmestik** (vordlusandmestik_synt_auto)
 - **Formaat:** Järgib eelmise andmestiku formaati, kuid süntaktilise märgendamise juures võib vastavalt muutuda ka lausefailide kuju. Metaandmefailis tehakse vajalikud täiendused.

- Moodustumine: Kasutatakse süntaksiparserit, näiteks UDpipe. Täpne valik toimub Tellija ja Täitja kokkuleppel, võttes aluseks pakutava parsimisväljundi sobivuse hanke eesmärkide tõhusaks ja küllaldaseks täitmiseks.
- Eesmärk: Luua süntaktilise märgendusega andmestik, mille aluselt läbi viia ühestamine.
- Millele tähelepanu juhtida: parseri väljundis nende kategooriate esinemine, mida kavatsetakse kasutada alusena hindamisaluste väljatöötamisel. Hinnanguline parseri toimetulek eesti õppijakeelega, et ühestamisprotsessi töömahtu võimalikult all hoida.

Tööde teostumise II järgus

- Süntaktilise käsitsimärgendusega võrdlusandmestik

(vordlusandmestik_synt_yhest)

- Formaati: Ühestatud kujul võrdlusandmestik, milles on kõrvutatavalt toodud parsimisvõimalused vigadega lausete, kõiki veaparandusi arvestavalt korrigeeritud lausete ja valikuliselt veaparandusi arvestavalt korrigeeritud lausete kohta. Iga korrigeeritava lause kohta, kui sellel on võimalikud ka paralleelsed analüüsivariandid, tuuakse välja ka need variandid (näiteks luues iga analüüsi kohta eraldi analüüsifaili). Vastavusviidete väljatoomiseks kasutatakse vastavat loogikat nii failinimedes kui ka metafailis.
- Moodustumine: Täitja määratud märgendajad kontrollivad Süntaktilise automaatmärgendusega võrdlusandmestiku käsitsi üle ja ühestavad selles olevad parsimisvead. Ühestatakse kategooriad: Lemma, Sõnaliik, Morfoloogiline kirjeldus, Süntaktiline kirjeldus, Sõltuvussuhte moodustajad ja liik. Lausepolüseemia korral ja sellest tuleneva erinevate parsimisvõimaluste olemasolul kirjeldatakse ka alternatiivsed süntaktilised märgendused. Eelistatavalt läbib kogu materjal topeltmärgenduse (käib läbi vähemalt kahe märgendaja käe alt).
- Eesmärk: Tekitada korpus, mille alusel oleks võimalik hinnata kas automaatparseri pakutav väljund õppijakeele materjalide analüüsimisel kuulub korrektsete analüüsivariantide hulka. Kui ei kuulu, siis võimaldada tuvastada, millised on eksimused ja kuidas need suhestuvad vigadega ja korrigeeritud teksti parsimisel, seejuures võimaldades juurdeloodava metoodika alusel ka tuletada, milliselt on parsimissooritus mõjutatud vigadega lauses keelevigade esinemisest/mitteesinemisest.
- Millele tähelepanu juhtida:

- Metoodika

- Formaat: Täpne formaat on täitja otsustada. Peab täitma uurimistöö läbiviimisele ja tulemuste selgele vormistamisele esitatavaid üldnõudeid. Metoodiliste lähtekohtade valikul toetutakse mh Huang jt (2018) uurimistööle.
- Moodustumine: Uurimistöö viib läbi Täitjapoolne uurija või uurijad.
- Eesmärk: Tekitada metoodilised alused, millega oleks võimalik Süntaktilise käsitsimärgendusega võrdlusandmestikku rakendades tõhusalt erinevate süntaksiparserite sooritust õppijamaterjali töötlemisel hinnata ja benchmarkida. Otsestelt rakendatavate benchmarkide väljatöötamist ei eeldata, piisab metoodiliste aluste väljatöötamisest.
- Millele tähelepanu juhtida: Täpsemad lahendused arutatakse läbi ja kooskõlastatakse projekti jooksul.
- **Terviklik üleantav andmestik, kood ja metoodika**
 - Hanke tulemusena valmib ja tellijale antakse terviklikuna üle:
 - Täitja koostatud kood Alusandmestikust Süntaktilise märgendusega võrdlusandmestiku genereerimiseks. Täitja tagab koodile piisava dokumentatsiooni ja migratsiooni Tellija koodirepositooriumisse.
 - Terviklik andmestik, mis koosneb eelpool kirjeldatud alaandmestikest (kõigi loodavate andmestike puhul järgitakse samalaadset struktuuriloogikat):
 - Alusandmestik (Tellija poolt)
 - Süntaktilise märgendusega võrdlusandmestik (vordlusandmestik_synt_ilma)
 - Süntaktilise automaatmärgendusega võrdlusandmestik (vordlusandmestik_synt_auto)
 - Süntaktilise käsitsimärgendusega võrdlusandmestik (vordlusandmestik_synt_yhest)
 - Hindamismetoodika

Hanke käigus Täitja kasutusse antavate materjalide sisu osas kehtib konfidentsiaalsuskohustus ning teavet ei tohi Täitja avaldada ilma õigusliku aluseta kolmandatele osapooltele. Töö lõppedes tagastab Täitja Tellijale töö käigus saadud materjalid koos kõigi teiste töö tulemusena valminud üleantavate materjalidega. Täitja endale materjalide koopiaid ei jäta. Hankevõitjaga sõlmitakse täiendav kokkulepe materjalide töötlemise, hoiustamise jms osas.

Täitja tagab, et tema poolt loodud või tema poolt kolmandatelt isikutelt omandatud mistahes tööde resultaadid (edaspidi materjalid) ja nendega seotud intellektuaalse omandi õigused (edaspidi õigused), sh autori kõik varalised õigused, lähevad materjalide vastuvõtmisega täies mahus üle tellijale ning materjalide suhtes kehtivate

autori isiklike õiguste osas (mis oma olemuselt üleantavad ei ole) annab töövõtja tellijale arvates materjalide üleandmisest tagasivõtmatu, kogu autoriõiguste kehtivuse aja kehtiva lihtlitsentsi ja õiguse anda all-litsentse kolmandatele isikutele enda äranägemisel (edaspidi koos nimetatud litsents)

3. Pakkujale ja pakkumusele esitatavad nõuded

- Pakkuja esitab pakkumises
 - Projektiplaani
 - Projektijuhi ja programmeeri CV ja kinnitus rollile vastavuse kohta.
- Pakkuja meeskonnas peavad olema tagatud järgmised rollid:
 - Projektijuht: Doktorikraadi omandanud või seda omandamas isik. Doktori- või magistrikraad peab olema keeleteaduses või sellele lähedasel erialal. Varasem (viimase 48 kuu jooksul) peab olema kogemus süntaktilist ühestamist puudutavate projektide juhtimise või täitmisega.
 - Programmeeri: IT-alase kõrg- või rakenduskõrgharidusega (või seda hetkel omandav) inimene, kellel on tõestatav varasem arendaja rolli täitmise kogemus vähemalt kahe keeletehnoloogiat puudutava projekti juures. Projektides töötamise kogemuse alternatiivina võib olla ka vähemalt 24 kuud arendajana töötamise töökogemust (samast vastates eelmainitud hariduskriteeriumidele).
 - Projektijuhi valikul märgendaja(d): Vähemalt magistrant. Omandatud või omandatav magistrikraad (või kõrgem kraad) peab olema keeleteaduses või sellele lähedasel erialal.
- Üks inimene võib täita ka mitut rolli (näiteks Projektijuht võib täita ka Programmeeri ja/või märgendaja rolli), kuid eraldi inimesed peavad olema vähemalt kahele rollile.
- **Projektiplan**, milles on kirjeldatud hanke läbiviimise tööplan, pakkumuskutses kirjeldatud andmestike loomist ja metoodika väljatöötamist puudutavate eesmärkide saavutamise kirjeldus. Esitatavas projektiplanis peab kirjelduma järgmine:
 - Hanke jooksul teostatavate tegevuste sisu ja ajaline jaotus. Kirjeldus antakse hanke kummagi etapi vaates. Tegevused on esitatud kindla algus- ja lõpptähtajaga (võttes aluseks nädalates kestvuse alates etapi perioodi algusest). Ülevaade esitatakse Gantti graafikul.
 - Materjalide süntaktiliseks parsimiseks (Võrdlusandmestik_koos_sünt andmestiku loomiseks) kasutatava parseri eelistus ja selle valimise põhjendus.
 - Ühestamise läbiviimise metoodilised ja korralduslikud alused. Selles on kirjeldatud, mille osas ühestatakse, milliseid failivorminguid kasutatakse,

mitme märgendajaga töö toimub ja kuidas lahendatakse erinevused märgendajate vahel.

- Mh tuuakse välja, kuidas kavatakse toimida erinevaid analüüsivõimalusi (parsimisvõimalusi) pakkuvate lausete ühestamisel. Tuuakse näiteid kolmest sellisest lausest ja erinevates ConLL-U (või võrreldavas failiformaadis) failides kirjeldatakse nende erinevad analüüsivõimalused.
- Hindamis- ja benchmarkimismetoodika väljatöötamise esmased uurimisküsimused ja lahendatavad probleemid.

Esitatava projektiplaani alusel hinnatakse pakkuja suutlikkust projekti teostada ja hanke eesmärgi saavutada. Kui tellija hindab esitatud projektiplaani sellele esitatud nõuetele mittevastavaks, siis on tal õigus esitatud pakkumine tagasi lükata.

Esitatud projektiplaani arutatakse läbi ja kooskõlastatakse tööde algusetapis. Täpsustatud ja kooskõlastatud projektiplaani võetakse tööde teostamise aluseks, kuid kahepoolsele kokkuleppele võib sellesse projekti teostamise vältel teha ka jooksvalt muudatusi.

4. Pakkumise tähtaeg

- 3. november 2025
- Pakkumisperioodil on pakkujal õigus pöörduda hankija poole täpsustavate küsimustega ning pärida võimalust aluskorpusse tutvumiseks.

5. Tööde perioodid valmimise tähtajad ja tööde eest tasumine

- **Töö valmib kahes etapis:**
 - **I etapp**
 - **Teostatavad tööd ja tulemus:** Esimese etapi jooksul valmivad Kood, Süntaktilise märgendusega võrdlusandmestik ja Süntaktilise automaatmärgendusega võrdlusandmestik. Etapi lõpus antakse see osa töödest tellijale üle.
Väljamakse maht: 40% hanke kogumaksumusest.
Perioodi kestvus: Perioodi tööde valmimise tähtaeg on neli nädalat hankelepingu sõlmimisest.
 - **II etapp**
 - **Teostatavad tööd ja tulemus:** Teise etapi jooksul valmivad Süntaktilise käsitsimärgendusega võrdlusandmestik ning benchmarkimis- ja hindamismetoodika. Etapi lõpus antakse Tellijale üle terviklik andmestik, mis koosneb töö käigus koostatud alaandmestikest.

Väljamakse maht: Tööde üleandmisel kuulub väljamaksmisele järelejäänud 60% hanke maksumusest.

Perioodi kestvus: Perioodi tööde valmimise tähtaeg on 12 nädalat esimese etapi tööde üleandmise tähtajast (17 nädalat hankelepingu sõlmimisest).

○ **Tasumise põhimõtted:**

Ettemaksu ei tehta.

Tasumine iga etapi lõpus pärast tööde üleandmise-vastuvõtmise akti vormistamist. Teostaja esitab arve hiljemalt 7 päeva jooksul pärast üleandmise-vastuvõtmise akti allkirjastamist. Maksetähtaeg arvel: 7 päeva.

6. Eduka pakkumuse valiku kriteerium

- Hanke maksimaalne eeldatav maksumus on 29 900.- (ilma käibemaksuta).
- Kaalutakse ainult nende pakkumiste vahel, mille Tellija tunnistab projektiplaani ja pakkuja meeskonna alusel tingimustele vastavaks. Mittevastavaks tunnistamise korral Tellija annab tagasisidet otsuse põhjuste kohta.
- **Madalaim hind**
- Pakkumus tuleb esitada e-posti aadressile kristjan.suluste@eki.ee hiljemalt ülaltoodud tähtajaks. Kõigi pakkujatega võetakse ühendust ja tehakse tulemus teatavaks hiljemalt 17. november 2025.
- Samaväärsete pakkumiste korral tehakse valik liisuheitmise meetodil.

7. Viited

- Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English.
<https://doi.org/10.17863/CAM.23072>