Error Classification & Annotation of Learner Language

for Developing Estonian Grammar Correction

Kais Allkivi-Metsoja,
Karina Kert, Silvia Maine,
Kaisa Norak, Pille Eslon

Tallinn University

Background

- Error annotation (error tagging) – marking language errors in a text corpus
  - Indicating error scope, type and correction
- Error annotated subset of Estonian learner language in development
  - Nationally funded project “Automated correction of Estonian language texts”
  - In collaboration with the University of Tartu, 2021–2023

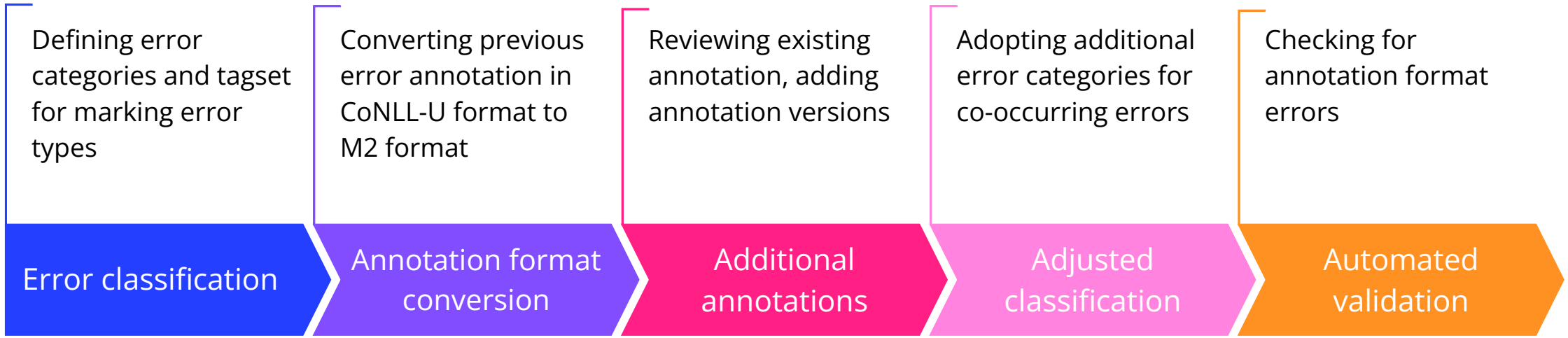
Need for error annotated language data

- Testing the performance (precision and recall) of grammatical error correction (GEC) systems
- Development of automated error annotation systems (e.g., ERRANT)
- Research on learners' language production, development of learning materials and tools

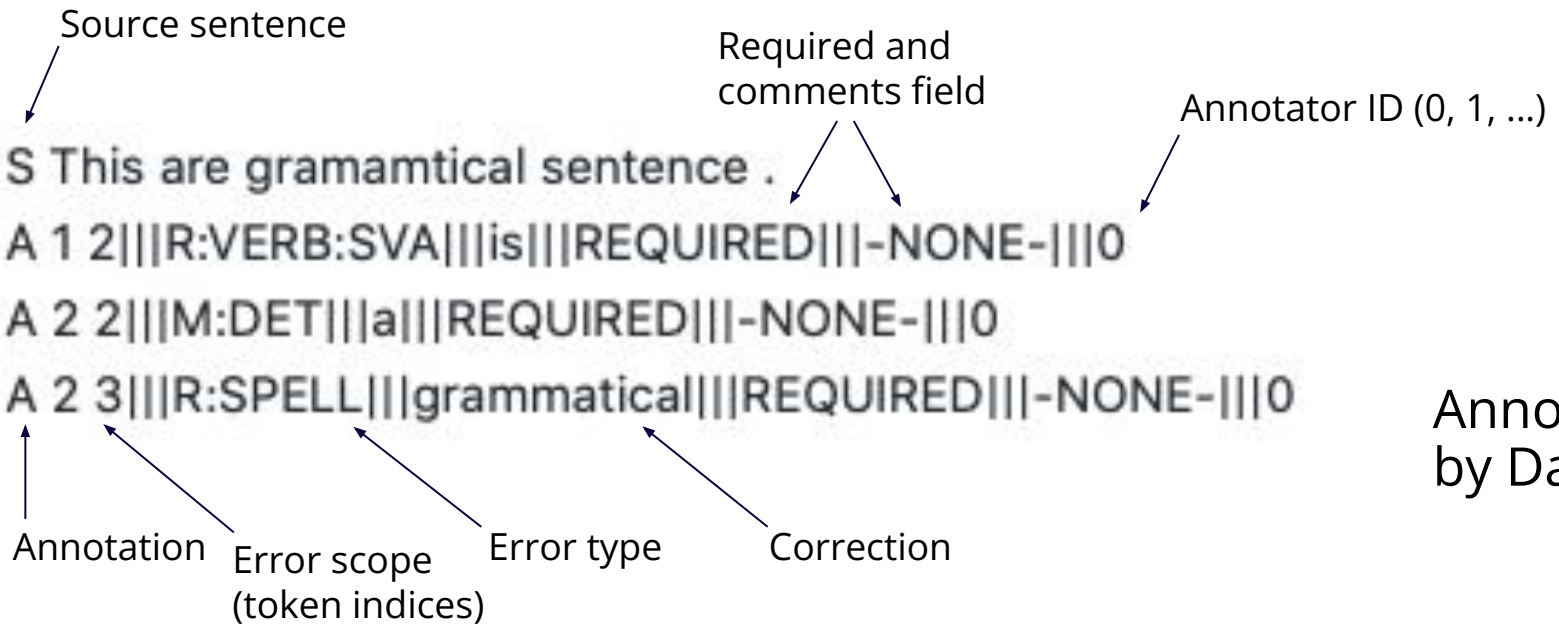
Corpus for testing Estonian GEC

- Estonian proficiency examination writings (level A2–C1) from the Estonian Interlanguage Corpus
- Initially annotated in the CoNLL-U format: error tags were added to morphologically tagged texts
- Converted to M2 format allowing several annotations
- Planned volume ~3,700 sentences
- ~2,500 sentences error annotated in M2 format: A2 – 495, B1 – 673, B2 – 723, C1 – 566

Timeline


Defining error categories and tagset for marking error types
Converting previous error annotation in CoNLL-U format to M2 format
Reviewing existing annotation, adding annotation versions
Adopting additional error categories for co-occurring errors
Checking for annotation format errors
Error classification
Annotation format conversion
Additional annotations
Adjusted classification
Automated validation

Error annotation in M2 format



Annotation format proposed by Dahlmeier & Ng (2012).

Main error classification

Replacement

- Spelling - R:SPELL
- Inflectional form - R:NOM:FORM, R:VERB:FORM
- Capitalization - R:CASE
- Whitespace - R:WS
- Word choice - R:LEX
- Word order - R:WO
- Punctuation - R:PUNCT

Missing

- Word(s) - M:LEX
- Punctuation - M:PUNCT

Unnecessary

- Word(s) - U:LEX
- Punctuation - U:PUNCT

Adapted from the ERRANT error classification framework for English (Bryant et al., 2017).

Differences to English error annotation

- Overlapping error scope in case of word order errors
  - Problematic for GEC testing tools developed for English
- Co-occurring error types -> compound error types
- Error type always defined, the tag “R:OTHER” not used

Preserving the source sentence

Ex 1. The grammatical errors in the source sentence are corrected, but the errors regarding style were neglected.

S Mulle meeldib see auto , aga ma ostsin uut autot .
Mulle meeldib see auto , aga ma ostsin uue auto .
‘I like this car , but I bought a new car .’
A 8 9|||R:NOM:FORM|||uue|||REQUIRED|||-NONE-|||0
A 9 10|||R:NOM:FORM|||auto|||REQUIRED|||-NONE-|||0

S Pärast kooli lõpetamist mina teda ei näinud .
A -1 -1|||noop|||-NONE-|||-NONE-|||-NONE-|||0
Pärast kooli lõpetamist ma teda ei näinud.
‘I didn’t see him after graduation.’
A 3 4|||R:NOM:FORM|||ma|||REQUIRED|||-NONE-|||1

Lexical choices

Ex 2. Inadequate word used in the source sentence.

S Ta oli ebanormaalne õpilane .
Ta oli ebatavaline||ebaharilik õpilane .
‘He was an abnormal student.’
A 2 3|||R:LEX|||ebatavaline||ebaharilik|||REQUIRED|||-NONE-|||0

Overlapping errors – compound error types

Ex 3. There are multiple errors in one token, which were corrected by using a single new error type.

S See on pealinn Islandil .

See on Islandi pealinn .

‘This is Iceland’s capital.’

A 2 4|||R:WO:NOM:FORM|||Islandi pealinn|||REQUIRED|||-NONE-|||0

Overlapping errors – compound error types

Ex 4. There are multiple errors in one token, which were corrected by using a single new error type.

S Ma kohtusin minu vanad Sobrad .
Ma kohtusin oma vanade sõpradega .
‘I met my old friends .’
A 2 3|||R:LEX|||oma|||REQUIRED|||-NONE-|||0
A 3 4|||R:NOM:FORM|||vanade|||REQUIRED|||-NONE-...
A 4 5|||R:NOM:FORM:SPELL:CASE|||sõpradega|||...

Multiple correction possibilities

Ex 5. In the source sentence a grammatically incorrect case was used, which could be corrected in two different ways.

S Olime õppinud koolis 12 aastat ja pärast veel olime õppinud Tallinnas 5 aastat .
Oleme õppinud koolis 12 aastat ja pärast oleme veel õppinud Tallinnas 5 aastat .
'We have studied at school for 12 years and later we also studied in Tallinn for 5 years.'

A 0 1|||R:VERB:FORM|||Oleme|||REQUIRED|||-NONE-|||0
A 3 3|||M:LEX|||koos|||REQUIRED|||-NONE-|||0
A 8 9|||R:VERB:FORM|||oleme|||REQUIRED|||-NONE-|||0
A 7 9|||R:WO|||oleme veel|||REQUIRED|||-NONE-|||0
A 10 13|||R:WO|||5 aastat Tallinnas|||REQUIRED|||-NONE-|||0

Õppisime koolis 12 aastat ja pärast õppisime Tallinnas veel 5 aastat.
'We studied at school for 12 years and later studied in Tallinn for another 5 years.'

A 0 2|||R:VERB:FORM|||Õppisime|||REQUIRED|||-NONE-|||1
A 8 10|||R:VERB:FORM|||õppisime|||REQUIRED|||-NONE-|||1
A 7 11|||R:WO|||õppisime Tallinnas veel|||REQUIRED|||-NONE-|||1

Annotation format checker

Examples of errors found using the annotation format checker

The length of annotation line 59 is incorrect:
A 6 8|||hotellis oli|||REQUIRED|||-NONE-|||0

Annotation line 9 has an unrecognisable error tag:
A 1 4|||R:W0|||olen ammu lemmikloomast unistanud|||REQUIRED|||-NONE-|||1

There is an extra space in annotation line 6:
A 4 6|||R:WO||| üldse loomi|||REQUIRED|||-NONE-|||0

THANK YOU!

TALLINN UNIVERSITY

Language Technology Research Group,
School of Digital Technologies

Follow our progress:

https://github.com/tlu-dt-nlp

Contact:

kais@tlu.ee