DARIAH-EU
CLARIN
Common Language Resources and Technology Infrastructure

FIN-CLARIAH

Ari Huhta
University of Jyväskylä
Venla Räsänen
University of Jyväskylä
Tanja Välisalo
University of Jyväskylä
Sirkku Lesonen
University of Jyväskylä

WP2.5 SOLUTIONS FOR BETTER USE OF LANGUAGE LEARNER PERFORMANCES IN RESEARCH

Learner language expressions are often non-conventionalized.
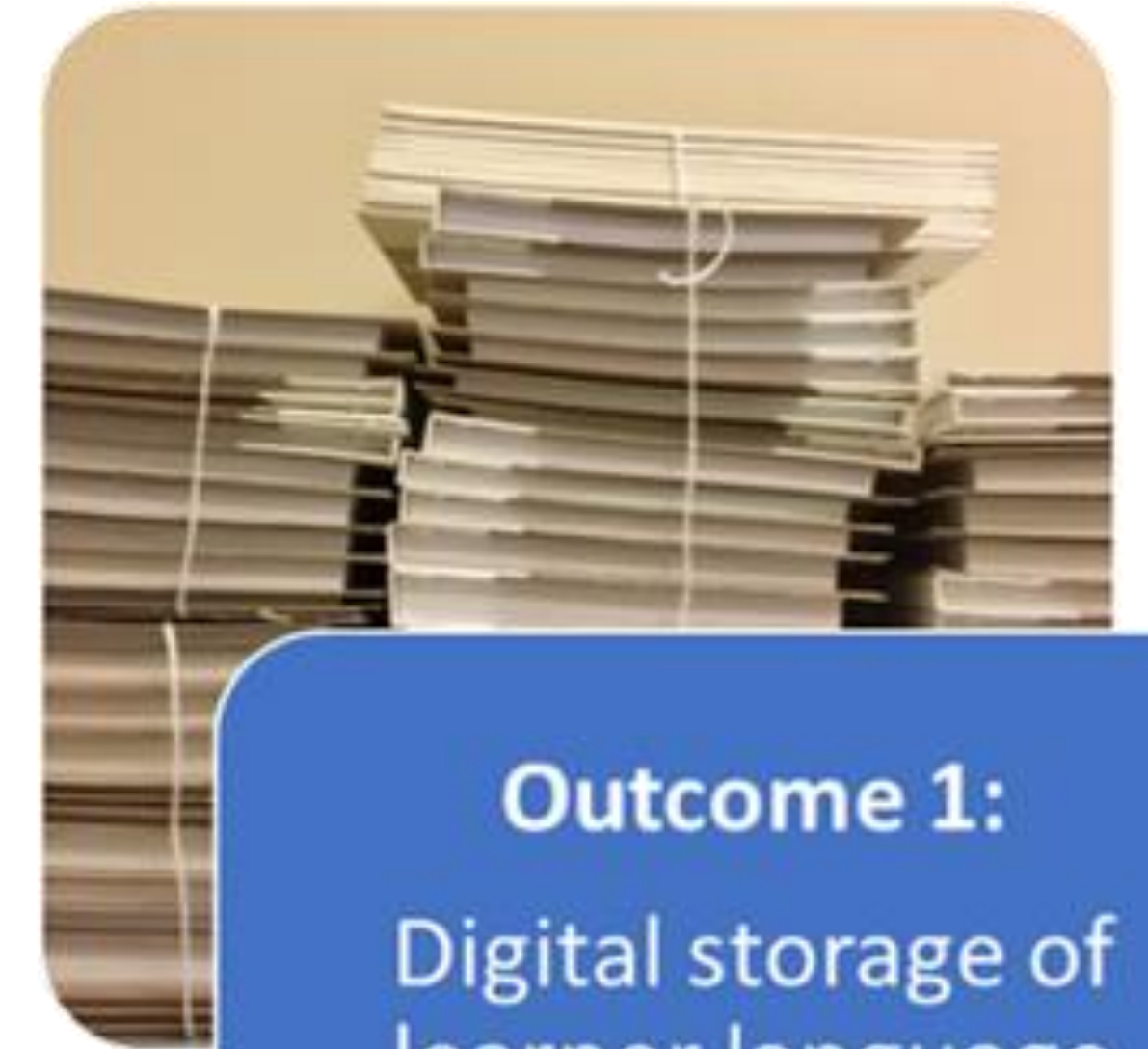
The school day shouldn't begin before 10 o'clock because the school dayes are very long in the afternoon. and because if school days finnished at four o'clock you are at home half past four and you doesn't have much time do everything.

Kännykät ei saa paista koulusta koska jos vaikka vanhemmille on jottain asia että sano lapsille hänen tarvitse saita hänelle ja jos paistetaan kännykät koulusta ne ei pysty saitaman tai lähettämän oma tärken asia.

Automated annotation and analysis of non-conventionalized learner language is challenging. However, there is a need for AI-based solutions so that learner language data could be used more efficiently in research in the fields of social sciences and humanities.
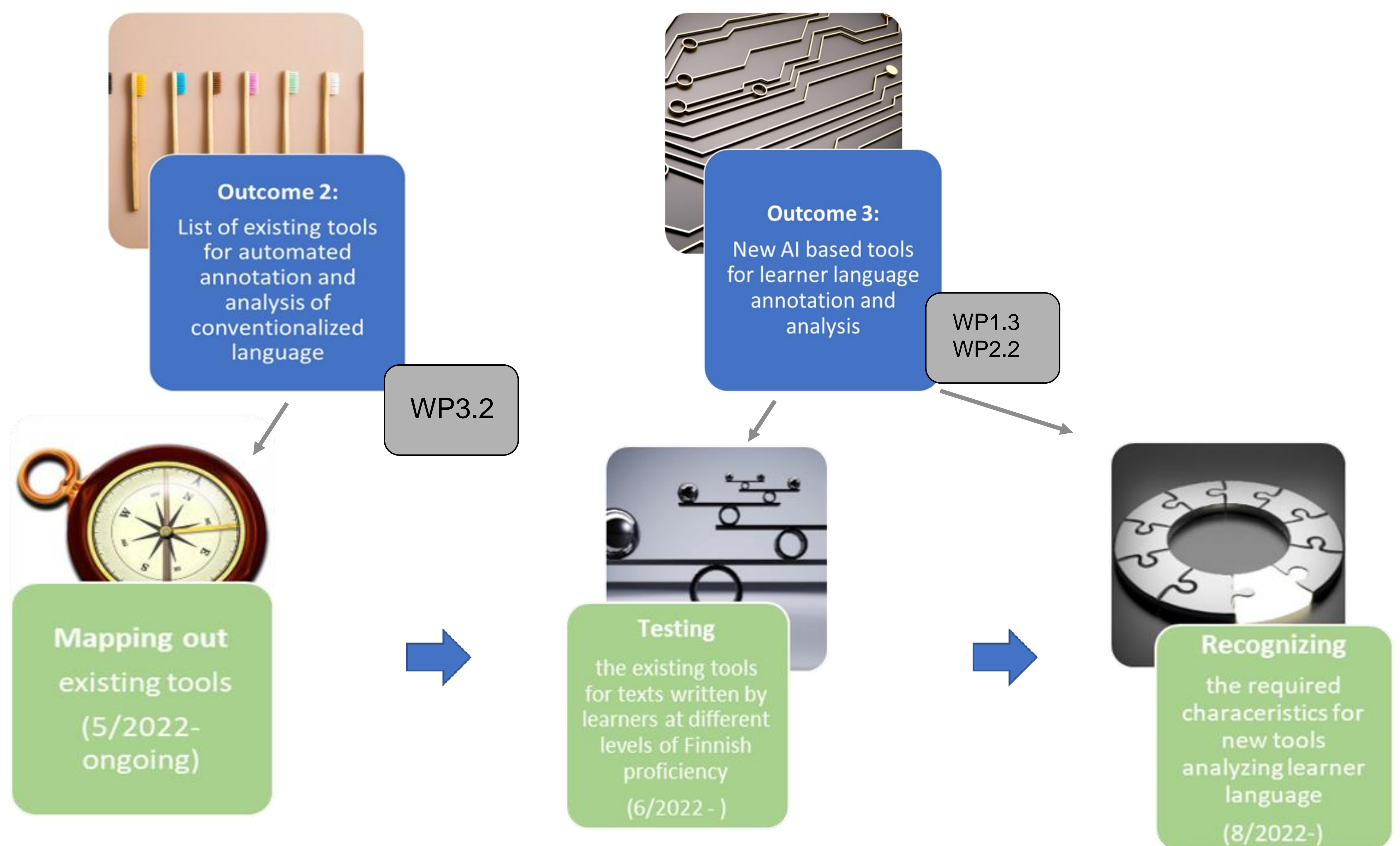
AIM

The aim of the WP2.5 is to provide infrastructure for a better use of language learner performances in research. Our aims are to 1) make public a large dataset of texts produced by Finnish language learners and 2) develop AI based tools for learner language annotation and analysis.

Outcome 1:
Digital storage of learner language texts produced for the National Certificates of Language Proficiency

OUTCOMES AND ROADMAP

Outcome 2:
List of existing tools for automated annotation and analysis of conventionalized language

WP3.2

Outcome 3:
New AI based tools for learner language annotation and analysis

WP1.3
WP2.2

Mapping out existing tools
(5/2022-ongoing)

Testing
the existing tools for texts written by learners at different levels of Finnish proficiency
(6/2022 -)

Recognizing
the required characeristics for new tools analyzing learner language
(8/2022-)

CHALLENGES

What challenges does low-proficiency Finnish language pose for textual analysis tools? Which aspects of learner language cause more vs less problems (content / meaning / vocabulary; structures / grammatical forms; textual features (e.g. coherence/cohesion))?