

# Master Thesis title / MAL izenburua

Deep understanding of CEFR guidelines for multilingual readability assessment

## Proposer(s) / Proposatzailea(k): names / izenak

Itziar Gonzalez-Dios and Kepa Bengoetxea

## Contact / Kontaktua: email

itziar@gonzalezd@ehu.eus

## Description / Deskribapena

The Common European Framework of Reference for Languages (CEFR) is the framework for classifying language proficiency in Europe. In this project, we propose to analyse these guidelines and corpus (exams, available learner corpora) of the different levels with the aim of determining which are the linguistic measures and vocabulary that is needed at each level. For example, in A1 level the requirement of the listening understanding is:

*"I can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly".*

In this case, regarding the lexical part, the student should propose clusters of words that are related to the family (father, mother...), the words concerning itself (tallness, colour of the hair...)... using e.g. wordnet or similarity algorithms.

The aim is to use these linguistic facts as features in a multilingual readability assessment.

## Goals / Helburuak

- 1) Determine linguistic evidences for CEFR
- 2) Improve the readability assessment beyond statistical models
- 3) Understand the complexity required for each CERF level in a multilingual environment

## Requirements / Betebeharrak

Linguistic background, language proficiency in at least three languages

## Framework / Esparrua

Readability assessment

## Tasks and plan / Atazak eta plana

- 1) Collect and understand the information of the CEFR guidelines
- 2) Contrast the the information of the CEFR guidelines to corpora in at least three different languages
- 3) Propose linguistic measures for certain levels