



An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data

Katrin Wisniewski², Karin Schöne², Lionel Nicolas⁴, Chiara Vettori⁴,
Adriane Boyd¹, Detmar Meurers¹, Andrea Abel⁴, Jirka Hana³

1. Department of Linguistics, Universität Tübingen, Germany.
2. Department of Romance Languages, Technical University Dresden.
3. Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
4. Institute for Specialised Communication and Multilingualism, European Academy of Bolzano, Italy.



The project in very short

What kind of project?

=> EU Lifelong Learning Programme (project number 518989-LLP-1-2011-1-DE-KA2-KA2MP)

Who?

=> 4 research institutions from Germany, Italy and Czech Republic.

Context and general aims?

=> Though the Common European Framework of Reference for Languages (CEFR) nowadays holds the leading role in language teaching and certification, it is often considered insufficiently illustrated in terms of authentic learner data (Fulcher 2004; Hulstijn 2007). MERLIN is designed to improve this situation.

Main objectives?

=> To research and enhance the empirical foundations of the CEFR scales by constructing a written learner corpus for German, Italian, and Czech as L2.
The annotated corpus and related information will be available to users on a freely accessible, didactically oriented online platform (see website).

Data Selection and Preparation

- => Corpus compiled using current best practices in data selection and preparation.
- => Standardized, CEFR-related tests of L2 German, Italian and Czech.
- => Tests compliant with the criteria of the Association of Language Testing in Europe.
- => Roughly 200 texts per examination level (German: A1–C1, Italian A1–B2, Czech A2–B2).
- => High degree of rating reliability on test analyses, including Multi-Facet Rasch analyses.

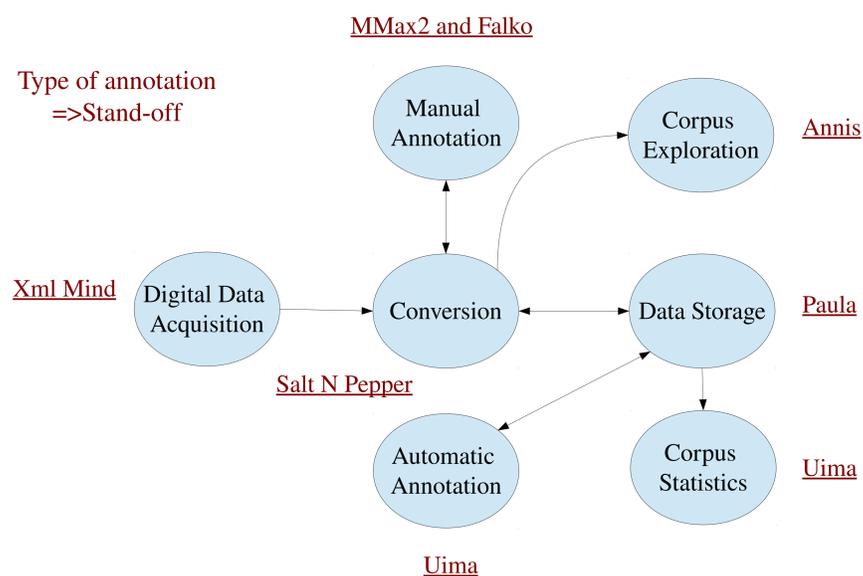
The Annotation Scheme

Annotations originating from various sources.

- (1) Tags designed to determine whether the predictions of the CEFR scales correspond to learner behavior => Selected CEFR scales were operationalized.
- (2) Research-based annotation of coherence, grammar, vocabulary, orthography, and sociolinguistics.
- (3) Questionnaire study and expert interviews (teachers and other envisaged user groups).
- (4) Textbook and language test analyses induced annotation schemes.
- (5) Inductive analyses were carried out to identify additional language features.

- => The sources of all tags will be transparently documented for MERLIN users.
- => Target hypotheses are provided for each learner production to explicitly record the forms on which the annotated interpretations are based.

Technical Infrastructure



“Some” of the Applications

MERLIN as illustration of CEFR levels

- * Originally devised to be general, the CEFR levels need supplementary language-specific illustrations.
- * Since 2001 the Council of Europe itself has encouraged the development and the circulation of accompanying tools which better illustrate the features of one single language.
- => MERLIN aims at illustrating CEFR levels as other projects but addresses three languages, supporting cross-language comparisons.
- => MERLIN provides the full texts, test tasks, and a wide range of linguistic and error annotations.

MERLIN for teaching, testing, and learning

- => The platform can help to identify and exemplify strengths and weaknesses of learners' performances at different levels and thus provide instructional progressions with a reliable grounding. With the help of the platform, users can observe and trace those features on the different CEFR levels to better understand how learners advance in their learning process.
- => The use of standardized MERLIN ratings will help revising personal or institutional assessment criteria and adjusting the rating behavior within a team of testers. The linguistic annotation enables users to run both basic and sophisticated searches on the basis of the full range of properties provided by the MERLIN Annotation Scheme.

MERLIN for the validation of the CEFR scales

- * Empirical validity of the CEFR scales has been researched little.
- * There is a well-known tendency of trained raters to base their decisions on scale-external aspects.
- * Looking for empirical language features that are typical of rated CEFR levels is not sufficient for examining empirical scale validity.
- => By including operationalized CEFR level descriptions in its Annotation Scheme, MERLIN allows to search for scale content correlates in learner language as directly as possible.

MERLIN to advance NLP for learner language

- => The corpus and its meta-information on learners and ratings readily supports research on automatic native language identification.
- => The corpus has already been used for research on automatic proficiency classification for German (Hancke 2013).
- => The corpus provides richly annotated learner data for the development and adaptation of NLP tools and applications assisting language learners in improving their vocabulary usage, coherence, spelling and grammatical accuracy.