The MERLIN Corpus

Learner language and the CEFR

Adriane Boyd[1], Jirka Hana[3], Lionel Nicolas[4], Detmar Meurers[1], Katrin Wisniewski[2], Andrea Abel[4], Karin Schöne[2], Barbora Štindlová[3], Chiara Vettori[4]

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.

Abstract

The MERLIN corpus is a written learner corpus for Czech, German, and Italian that has been designed to illustrate the Common European Framework of Reference for Languages (CEFR) with authentic learner data. The corpus contains 2,290 learner texts produced in standardized language certifications covering CEFR levels A1--C1. The MERLIN annotation scheme includes a wide range of language characteristics that enable research into the empirical foundations of the CEFR scales and provides language teachers, test developers, and Second Language Acquisition researchers with concrete examples of learner performance and progress across multiple proficiency levels. For computational linguistics, it provide a range of authentic learner data for three target languages, supporting a broadening of the scope of research in areas such as automatic proficiency classification or native language identification. The annotated corpus and related information will be freely available as a corpus resource and through a freely accessible, didactically-oriented online platform.

Motivation

(1) Validation of the CEFR Scales

x CEFR scales used for multiple purposes, even high-stakes tests,
x CEFR scales often considered as insufficiently illustrated (consensual doubts about them actually mirroring what learners do).
=> The corpus illustrates CEFR-based ratings and contributes to a rating scale validation.
=> The relationship between CEFR scales and empirical learner language is studied by integrating CEFR level descriptions in the annotation scheme.

(2) Teaching, Learning, and Testing

x Learner corpora had so far little impact on teaching. As it is publicly available and one of only a few corpora related to the CEFR, the MERLIN corpus directly addresses such issue.
=> The use of standardized MERLIN ratings helps revising assessment criteria.
=> The large set of linguistic annotation enables users to run sophisticated searches.
=> The platform can help observing learners' performances with a reliable grounding.
=> Features on different CEFR levels allow to observe the learning process of learners.

(3) NLP for Learner Language

The corpus provides valuable data for natural language processing of learner language.
=> Automatic leaner proficiency classification and native material readability analysis,
=> Automatic native language identification (beyond the focus on English learners)
=> Richly annotated learner use-case data for developing NLP tools to assist learners.

The Annotation Scheme

Annotation scheme

Various sources considered for establishing the annotation scheme:
(1) CEFR scales operationalisation for studying/matching scales with learner behavior.
(2) SLA and language testing research (orthography, grammar, vocabulary, coherence & cohesion, sociolinguistic appropriateness & pragmatics)
(3) Questionnaire study and expert interviews (teachers and other envisaged user groups).
(4) Experientially derived indicators from textbooks and empirical analyses of samples.

Annotation scheme Design

=> 2 types of annotations focusing on errors and noteworthy use of certain structures.
=> 3 levels: linguistic field, linguistic subfield and a third, optional, sub- specification.
=> 7 linguistic fields ranging from orthography to pragmatics.

Target Hypotheses

=> Explicit records of annotators' interpretations to ensure coherence between annotations.
=> 2 types of target hypotheses performed following the guidelines of the FALKO project:
minimal target hypothesis (TH1), focusing on linguistic correctness,
extended target hypothesis (TH2), focusing on linguistic appropriateness.

Learner Text

Tokens	Ich	möchte	mich	bei	Ihnen	um	eine	vertriebspraktikantenstelle	im	Bereit	als	IT-Systemkaufmann	bewerben	.
Gloss	*I*	*would like*	*myself*	*to*	*you*	*for*	*a*	*sales trainee position*	*in the*	*field*	*as*	*IT-systems assistant*	*apply*	.
Translation	'I would like to apply for a sales trainee position in the field of IT-systems assistant.'													

Target Hypothesis 1

TH1	Ich	möchte	mich	bei	Ihnen	um	eine	Vertriebspraktikantenstelle	im	Bereich		IT-Systemkaufmann	bewerben	.
TH1 Diff								CHA		CHA	DEL			

MERLIN Annotation

Orthography								Capitalization (error tag)		A[1]				
Grammar											B[2]			
Vocabulary			Collocation (existence tag)											
Socioling.	Opening/closing formula (existence tag)													
Coherence/Cohesion														
Language Functions														
Sentence Intelligibility	Completely intelligible (existence tag)													

[1]A – grapheme error: incorrect element (error tag)
[2]B – error in conjunctions: superfluous element (error tag)

Detailed German Annotation Example

Data

Selection and Preparation

=> Standardized, CEFR-related tests of L2 German, Italian and Czech
=> Tests audited by ALTE (Association of Language Testing in Europe).
=> ~200 texts per examination level (German: A1–C1, Italian A1–B2, Czech A2–B2).

Metadata

7 types of metadata:
=> learners' (1) age, (2) gender, and (3) L1,
=> (4) CEFR level of the test, (5) test institution, (6) test data, and (7) test task.

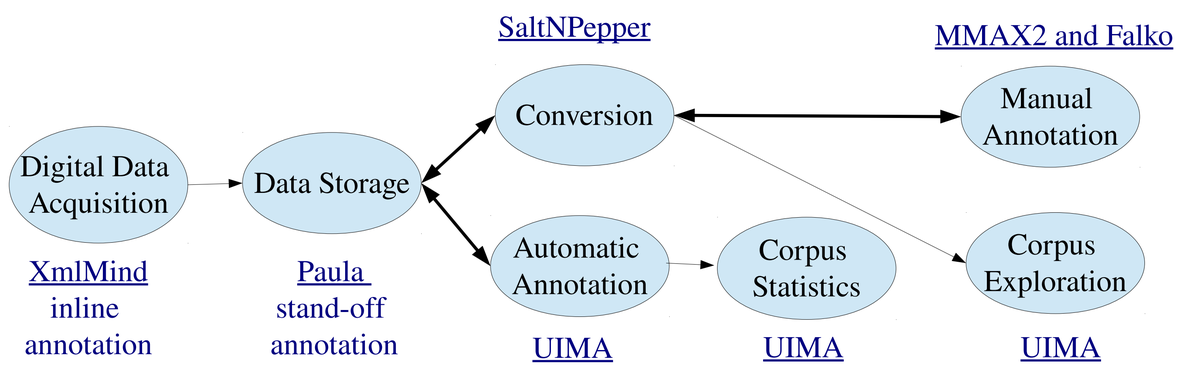
CEFR Ratings

=> Written learner productions extracted from the original tests
=> Rated accordingly to a CEFR-compliant analytical rating focusing on grammatical accuracy, vocabulary range & control, coherence/cohesion, orthographic control, and sociolinguistic appropriateness.
=> A holistic rating scale completed the grid and productions were assigned CEFR levels.
=> Analyses of rating (multi-facet Rasch analyses) showed a good reliability and allowed calculation of fair averages for balancing rater severity.

CEFR	A1	A2	A2+	B1	B1+	B2	B2+	C1	C2	Nb Texts	Nb Words
Czech	1	49	112	89	75	72	9	4		**411**	**64 488**
German	57	199	107	219	115	219	73	42	4	**1035**	**125 927**
Italian	30	294	94	343	53	2				**816**	**92 359**
Total	**88**	**542**	**313**	**651**	**243**	**293**	**82**	**46**	**4**	**2262**	**282 774**

MERLIN corpus by language and fair CEFR level

Annotation workflow



5 things to remember about the Merlin corpus

1) German, Italian, and Czech written learner corpus.
2) Designed to illustrate and validate the level system of the Common European Framework of Reference for Languages (CEFR) with authentic learner data.
3) Annotation scheme solidly grounded in a user-needs study, SLA research, inductive learner text analyses, and an operationalization of CEFR scales.
4) Provides learner metadata, detailed analytic and holistic CEFR ratings and a comprehensive set of learner and linguistic annotations.
5) Will be freely available at the end of 2014.