The MERLIN corpus: Learner language and the CEFR

Adriane Boyd,[1] Jirka Hana,[2] Lionel Nicolas,[3] Detmar Meurers,[1] Katrin Wisniewski,[4]
Andrea Abel,[3] Karin Schöne,[4] Barbora Štindlová,[2] Chiara Vettori[3]
[1]Universität Tübingen, [2]Charles University in Prague, MFF/FF, [3]EURAC, [4]TU Dresden
[1,4]Germany, [2]Czech Republic, [3]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 provide 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.

Keywords: learner corpora, corpus annotation, language proficiency, language testing, CEFR, Czech, German, Italian

1. Introduction

While the Common European Framework of Reference for Languages (CEFR) holds a leading role in language teaching and certification, there are serious concerns about the lack of a systematic empirical illustration of the CEFR centerpiece, its common reference levels. These concerns grow even stronger when languages other than English are considered (Fulcher, 2004; Hulstijn, 2007). The EU Lifelong Learning Programme project MERLIN "Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context" (2012–2014) aims at addressing these concerns (Wisniewski et al., 2013). In this paper, we present the MERLIN corpus, a written learner corpus for Czech, German, and Italian as a second language that has been devised to illustrate the CEFR with authentic learner data and support research into and enhancement of the empirical foundations of the CEFR scales.

The MERLIN corpus contains learner texts along with metadata about the learners and the test situation in which the texts were produced. It includes manual and automatic annotation for a wide range of language characteristics. Complementing the goal of supporting research into the empirical foundations of the CEFR scales and Second Language Acquisition, the corpus also supports research at the interface of computational linguistics and language learning, such as automatic proficiency classification or native language identification. Work in this area arguably has been limited by its almost exclusive focus on learners of English, so that the learner data for a Romance, a Slavic, and another Germanic language as provided by the MERLIN corpus will be particularly relevant for broadening the empirical basis of this emerging domain of research. The annotated corpus and related information will become freely available as a corpus resource and through a freely accessible, didactically-oriented online platform (http://www.merlin-platform.eu) at the conclusion of the project.

2. Motivation

Even though the descriptions of the CEFR levels needed to be general so as to be applicable across European languages (North, 2000), it was recognized that the descriptors would need supplementary language-specific illustrations. In that perspective, since 2001 the Council of Europe itself has encouraged the development and the circulation of accompanying tools which better illustrate the features of individual languages. One such initiative was the publication of the Reference Level Descriptions (RLDs) for national and regional languages.[1] Some initiatives increasingly base RLDs upon learner corpora, most prominently, the English Profile,[2] but also the Italian (Spinelli and Parizzi, 2010) and the Norwegian ones (Carlsen, 2013).

While being similar to these initiatives illustrating the CEFR levels for specific languages, the MERLIN corpus presents the advantage of addressing not one but three languages from different language families and supporting cross-language comparisons through a common annotation scheme. In addition, it distinguishes itself from related initiatives by providing free access to the full texts, test tasks, and a wide range of linguistic and error annotations.

The MERLIN corpus also contributes to several related domains of research presented in the following sections.

2.1. Validation of the CEFR Scales

The CEFR scales are used for a wide variety of purposes, among which are high-stakes tests that have serious consequences on the test-takers' lives. Nevertheless, evidence supporting the empirical validity of the CEFR scales are often considered as insufficient (Fulcher, 2004; Hulstijn,

This publication reports on work from the MERLIN project, funded by the European Commission (518989LLP1 2011DEKA2KA2MP). It only reflects the views of the authors and the Commission cannot be held responsible for any use which may be made of the information contained therein.

[1] http://www.coe.int/t/dg4/linguistic/dnr_en.asp

[2] http://englishprofile.org

2007). Even though they were calibrated in a multi-levelled procedure that included sophisticated statistical analyses, no learner language was analyzed in order to examine the potential of the scales to mirror what learners typically do in assessment situations (North, 2000).

Therefore, apart from illustrating CEFR-based ratings, MERLIN also contributes to a rating scale validation. In pursuit of that aim, the relationship between selected CEFR scales and empirical learner language can be analyzed as directly as possible by integrating operationalized CEFR level descriptions in the MERLIN annotation scheme (Wisniewski, 2013; 2014).

2.2. Teaching, Learning, and Testing

In spite of their potential, learner corpora so far have had relatively little impact on teaching (Gilquin et al., 2007). The MERLIN corpus, publicly available and one of only a few corpora related to the CEFR, is a particularly valuable resource in this context. The corpus can contribute to realistic expectations about learner achievements and inform the development of syllabi (Granger, 2009), allowing for an empirically validated selection of relevant language features that match learners' proficiency on all CEFR levels. Relatedly, the corpus can be helpful for standard-based development of teaching material. Textbooks or learner dictionaries can be related to the CEFR with MERLIN. CEFR-linked word lists are a powerful empirical complement to reference-level descriptions that are not corpus-based, such as Profile Deutsch (Glaboniat et al., 2003). More direct MERLIN applications in language teaching include opportunities for exploratory, data-driven autonomous learning (Gilquin and Granger, 2010). By offering access to tasks, task descriptions, ratings, full texts, annotations, and metadata of the learners, co-text and context are made transparent (Braun, 2005). Thus, learners, teachers, or testers can see how, for example, learners of the same L1 typically perform on different CEFR levels. The MERLIN project's user needs orientation, the lack of which is an often criticized corpus design criterion (Roemer, 2008), ensures the relevance of the annotated phenomena.

Despite Alderson's (1996) often cited argument for using more corpus information in language testing, learner corpora are not yet commonly referred to (Taylor and Barker, 2008). An exception is the Cambridge Learner Corpus (CLC) with the adherent English Profile Project[3] (Hawkins and Filipović, 2012), but the CLC unfortunately is not freely available for research.

The MERLIN corpus can be used as a continuous database for benchmarking and standard-setting, as a resource for rater training, and for validation of item writers' intuitions in the development of assessment material. It thereby helps relate language tests to the CEFR (Eckes, 2009) and to make cross-language comparisons of typical language characteristics on all CEFR levels. It should also make it possible to pursue the development of "criterial features" (Hawkins and Filipović, 2012) for German, Italian, and Czech. Overall, the platform is a powerful resource for empirically-based rating scale construction (Bachman and Palmer, 2010; Fulcher et al., 2011).

2.3. NLP for Learner Language

Complementing the uses directly linked to language teaching and testing, the MERLIN corpus also provides valuable data for the development and evaluation of natural language processing tools in this domain (Meurers, 2013).

The corpus has already been used for research on automatic proficiency classification for German (Hancke, 2013; Hancke and Meurers, 2013). The measures of second language proficiency employed in this context also are of relevance beyond the language learning context in that they have been shown to support the automatic analysis of the readability of native language material (Vajjala and Meurers, 2012; Hancke et al., 2012).

The corpus and its meta-information on the learners readily supports research on automatic native language identification (Tetreault et al., 2013). In line with Aharodnik et al. (2013), this will support such research beyond the currently dominant, narrow focus on English learners, which we expect to become particularly interesting in terms of the morphology and word order characteristics offered by the three target languages in the MERLIN corpus.

In the context of intelligent tutoring systems and writer's aids, the MERLIN corpus provides richly annotated learner data of relevance for the development and adaptation of NLP tools and applications that assist language learners in improving their vocabulary usage, coherence, spelling and grammatical accuracy.

3. Data

The MERLIN corpus was compiled from standardized, CEFR-related tests of L2 German, Italian (telc institute, Frankfurt) and Czech (ÚJOP Institute, Prague). The tests have undergone the strict auditing procedures of the Association of Language Testing in Europe (ALTE)[4] and comply with international test quality standards. The current corpus consists of about 200 texts per examination level for German A1–C1, Italian A1–B2, and Czech A2–B2.

3.1. Metadata

The corpus includes metadata about the learner's age, gender, and L1 along with information about the CEFR level of the test, the test institution, the test data, and the test task. For each CEFR test level and target language, there are at least two different tasks in the corpus. A typical task involves writing a letter to a friend or business in reply to a prompt with an advertisement, a letter, or a description of a personal situation. For example, the task for the corpus excerpt we will discuss below Figure 4 provides a job advertisement from a magazine and asks the learner to write a formal letter applying for the position.

3.2. CEFR Ratings

Written learner productions were extracted from the original tests and rated accordingly to a CEFR-compliant analytical rating grid by trained professional raters. This rating instrument is an assessor-oriented (Alderson, 1991) adaptation of the CEFR scales for grammatical accuracy, vocabulary range & control, coherence/cohesion, orthographic

[3] http://www.englishprofile.org

[4] http://www.alte.org

control, and sociolinguistic appropriateness and resembles Table 3 of the CEFR (Council of Europe, 2001). This grid is complemented by a holistic rating scale, the 'general linguistic range' of Council of Europe (2001). The learner productions were assigned CEFR levels for all of these rating criteria, resulting in individual competence profiles. Analyses of rating reliability (multi-facet Rasch analyses) were carried through and demonstrated a good degree of rating reliability (Wisniewski et al., 2013).

The Rasch analyses also allowed the calculation of fair averages for the holistic scale 'general linguistic range', which separate rater severity from learner proficiency (Eckes, 2009). Table 1 shows the distributions of texts by fair CEFR level, i.e., the level at which the learner performed as opposed to the level of the examination.

Fair CEFR Level		Czech	German	Italian	Total
Basic	A1	1	57	30	88
	A2	49	199	294	542
	A2+	112	107	94	313
Independent	B1	89	219	343	651
	B1+	75	115	53	243
	B2	72	219	2	293
	B2+	9	73		82
Proficient	C1	4	42		46
	C2		4		4
Number of Texts		411	1,035	816	2,262
Number of Words		64,488	125,927	92,359	282,774

Table 1: MERLIN corpus by language and fair CEFR level

4. Annotation Scheme

The MERLIN annotation scheme includes a wide range of language characteristics originating from various sources. The annotation scheme was developed to identify meaningful indicators that describe aspects of learner language from two perspectives. The first is more influenced by Foreign Language Teaching and Learning practice and annotates differences between the learner productions and the native target language norms (i.e., learner errors), whereas the second perspective is more in line with Second Language Acquisition (SLA) research, emphasizing the linguistic characteristics of the learner language.

The first perspective directly addresses differences between learner productions and correct second language utterances, i.e., error annotation. To support high-quality error annotation, the MERLIN corpus is annotated with explicit target hypotheses (Lüdeling, 2008) described here in section 4.3.

The second perspective is that of interlanguage as a dynamic and complex system. Learners' interlanguage depends on several factors including their native language, other languages they might know, the stage and ways of learning the language, etc. While interlanguage deviates from standard language, it is far from arbitrary. The texts can thus be annotated in the same way as any other texts with linguistic annotation. The view of learner language as an evolving language system in its own right is an important aspect of the MERLIN project and is reflected in the annotation scheme.

4.1. Annotation Scheme Development

The development of the annotation scheme incorporates features and characteristics from several, distinct sources:

Operationalization of the CEFR scales The annotation scheme includes tags that have been designed to determine if the CEFR scale contents can be transferred/found in learner language in order to run analyses of empirical scale validity (Fulcher, 2004; Hulstijn, 2007; Wisniewski, 2014). Selected CEFR scales were therefore operationalized (Wisniewski, 2013; Wisniewski, 2014). Such tags include intelligibility of the text, connector accuracy, content jumps, and collocation usage.

SLA and language testing research An extensive review of second language acquisition and language testing literature (cf., e.g., Carlsen (2010), Bulté and Housen (2012), Lu (2011), Malvern et al. (2004), Bestgen and Granger (2011), Trosborg (1995)) led to annotation tags in the areas of orthography, grammar, vocabulary, coherence/cohesion, and sociolinguistic appropriateness. The research-based annotations include:

orthography
grapheme-based errors, punctuation, capitalization, erroneous word boundaries, . . .

grammar
valency, agreement, word order, negation, . . .

vocabulary
different aspects of lexical knowledge with a particular focus on formulaic sequences, lexical errors, . . .

coherence/cohesion
connectors, use of text structural means, . . .

sociolinguistic appropriateness/pragmatics
addressing, requests, . . .

Teacher and expert interviews Thanks to a questionnaire study and expert interviews, teachers and other envisaged user groups indicated specific CEFR illustration needs for the MERLIN annotation scheme to cover. Accordingly, specific annotations have been designed such as verbal aspect in Italian and Czech and apostrophe use in German and Italian.

Experientially derived indicators Additional annotations, e.g., errors in double negation in Czech, have been directly derived from reference textbooks and language test analyses such as *Tangram* for German (Jan et al., 1998), *Rete!* for Italian (Mezzadri, 2000) and *Brána jazyka českého otevřená* for Czech (Hasil et al., 2007).

Inductive analyses of a sample of the texts (ten texts per CEFR level and language) produced empirically relevant learner language features such as article and clitic usage, the level of formality with respect to register, semantic errors, the use of formulaic sequences, citations from the test task or repetitions.

Linguistic Field	**Subfield**	**Phenomenon**
Orthography	Grapheme	Transposition* Accent*
	Word boundary	Split* Merge*
Grammar	Negation	Double negation*
	Verb	Tense* Voice*
Vocabulary	Formulaic sequence	Collocation Idiom
	Form	Deviation* Composition*
Coherence/Cohesion	Coherence	Text structural means
	Connectors	Accuracy*
Sociolinguistic Appropriateness	Letter text type	Greetings/farewells Opening/closing formulae
Pragmatics	Request	Direct request Indirect request
General	Text intelligibility Sentence intelligibility	

*Error tag

Figure 1: Excerpt from the MERLIN Annotation Scheme

4.2. Annotation Scheme Design

The MERLIN annotation scheme is organized into three hierarchical levels. Two obligatory levels describe the linguistic field and subfield. If needed, a third level further specifies the type of annotation. An excerpt of the annotation scheme is shown in Figure 1.

There are seven top-level linguistic fields ranging from orthography to pragmatics and a wide range of both error annotations and annotations for the learner's use of structures such as idioms and greetings/farewells in a letter.

4.3. Target Hypotheses

As multiple interpretations of a learner utterance are possible, it is necessary to explicitly record an annotator's interpretation of the learner's utterance in order to produce high-quality error annotation with high degree of interannotator agreement.

In the MERLIN corpus, these target hypotheses (Lüdeling et al., 2005) have been annotated for each learner production following the guidelines developed for the FALKO project (cf. Reznicek et al. (2012)). The FALKO annotation guidelines define two types of target hypotheses: a *minimal target hypothesis* (TH1) "should stay as close to the learner surface structure as possible" while an *extended target hypothesis* (TH2) "should reflect as much of the learner's intention in the utterance as possible" (Reznicek et al., 2013). The TH1 includes a minimal number of changes to the orthography, morphology, and syntax in order to produce a well-formed utterance in the target language. In contrast, the TH2 may include modifications to lexical, semantic, pragmatic, and stylistic aspects of the original text. The TH1 focuses on linguistic correctness, whereas the TH2 can be understood as focusing more on linguistic appropriateness (Reznicek et al., 2012).

A Czech example with both a minimal (TH1) and expanded (TH2) target hypothesis is shown in Figure 2. This sentence was written by a female learner in her 20s with L1 German in an A2 exam. The text received the fair CEFR level A2. In the TH1, three tokens have been modified due to orthographic errors. The tokens *něměčtínu* (correctly *němčinu* 'German$_{acc}$') and *Karlové* (*Karlově* 'Charles'$_{loc}$') both contain accent errors and the token *Univerzitě* (*univerzitě* 'university$_{loc}$') contains a capitalization error. In the TH2, the original token *něměčtínu* is modified further than in the TH1 to correct a vocabulary error in a derived word form.

When performing error annotation, annotators refer to the TH1 for orthography and grammar errors and to the TH2 for errors in vocabulary, coherence/cohesion, sociolinguistic appropriateness, and pragmatics. Each annotated error should correspond directly to one or more edits in the target hypothesis. See the following section for an example with a minimal target hypothesis and error annotation.

While the FALKO guidelines were developed for advanced learners of German, the MERLIN corpus contains many performances from beginning and intermediate learners, who are generally underrepresented in current learner corpus research as most corpora focus on the intermediate–advanced range (cf. Granger, 2008). As a result, some of the FALKO guidelines were adapted and new elements were added. The most significant addition is a new annotation layer where annotators indicate their uncertainty about how to formulate a target hypothesis. There are three levels of uncertainty: -1- is used when the intended meaning is uncertain, but inferable; -2- is a wild guess; and -3- indicates that no target hypothesis can be given as the sentence is not comprehensible. Figure 3 shows an example of an uncertain target hypothesis from an Italian text written by a 26-year-old female learner with L1 Hungarian. As the first word in the sentence is not a word in Italian and there are numerous edits required to form a target hypothesis, the annotator's uncertainty for the whole sentence is annotated in the "TH1 Uncertainty" layer.

Learner Text

Tokens	Je	profesorka	z	Německa	a	učí	němčtínu	na	Karlové	Univerzitě	v	Praze	.
Gloss	*is*	*professor*	*from*	*Germany*	*and*	*teaches*	*German*	*at*	*Charles*	*University*	*in*	*Prague*	.
Translation	‘She is a professor from Germany and she teaches German at Charles University in Prague.’												

Target Hypothesis 1

TH1	Je	profesorka	z	Německa	a	učí	němečtinu	na	Karlově	univerzitě	v	Praze	.
TH1 Diff							CHA		CHA	CHA			

Target Hypothesis 2

TH2	Je	profesorka	z	Německa	a	učí	němčinu	na	Karlově	univerzitě	v	Praze	.
TH2 Diff							CHA		CHA	CHA			

Figure 2: Minimal and Expanded Target Hypotheses

Learner Text

Tokens	Lo	sai	che	incontro	molti		,	chi	può	aiutarci	.
Gloss	*It*	*(you) know*	*that*	*(I) meet*	*many*			*who*	*can*	*help*	*us* .
Translation	‘You know that I meet many [people?] who can help us.’										

Target Hypothesis 1

TH1	Lo	sai	che	incontro	molti	individui		che	possono	aiutarci	.
TH1 Diff						INS	DEL	CHA	CHA		
TH1 Uncertainty	1 (=uncertain, but intended meaning can be inferred)										

Figure 3: Annotation of Uncertain Target Hypotheses

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)

Figure 4: Detailed Annotation Example

4.4. MERLIN Annotation Examples

Figure 4 shows the MERLIN annotation for a German learner sentence written in a B2-level exam by a female learner in her 30s. The full text received the fair CEFR level B1+.

The first section shows the tokens from the learner text along with an English gloss and translation. The second section contains the minimal target hypothesis (TH1), which provides the basis for the error annotation. The TH1 diff shows the individual token-level changes that were made to the sentence in the target hypothesis. The tokens *vertriebspraktikantenstelle* and *Bereit* ‘ready’ were modified (CHA) to the correct tokens *Vertriebspraktikantenstelle* and *Bereich* ‘area’ and the unnecessary token *als* ‘as’ was deleted (DEL).

The final section contains annotation from the MERLIN annotation scheme. The two CHA modifications in the target hypothesis correspond to the orthography error tags related to capitalization and the use of an incorrect grapheme (*t* instead of *ch*). The DEL modification is annotated with two errors, one in grammar (‘error in conjunctions: superfluous element’ and one in coherence/cohesion (‘usage of connectors: superfluous element’). Where appropriate, the error annotations specify whether there was a superfluous, missing, incorrect, or incorrectly ordered element.

The remaining annotations describe the presence of linguistic constructions and properties of the text rather than errors. In this sentence, the construction *sich bei jdm. um etw. bewerben* 'apply to so. for sth.' is annotated as a collocation and the entire sentence is part of an opening/closing formula. Finally, the sentence is annotated as being completely intelligible.

4.5. Automatic Annotation

The learner texts and target hypotheses are also annotated with automatic tools including part-of-speech taggers, lemmatizers, and parsers. The multi-layer standoff PAULA format (see details in the following section) makes it possible to easily extend the annotation with a new layer for any relevant tools. For example, the German subcorpus will include complexity measures used in Hancke (2013) to automatically predict learner proficiency from linguistic features in the learner texts.

5. Data Preparation and Annotation Workflow

The hand-written learner productions provided by the test institutions were scanned and then transcribed using XMLmind[5] with a custom stylesheet related to the text structure and digitalization process. Transcribers annotated insertions and deletions during the writing process along with ambiguous and unreadable tokens. They also anonymized all personal and place names with language-specific substitutions, identified foreign words and direct citations of the test prompts, and annotated greetings and closings for responses in letter format. Once transcription was completed, all data was converted to PAULA (Chiarcos et al., 2008), a standoff XML format that has be chosen for the distribution of the MERLIN corpus.

Manual annotations are being performed with two tools: MMAX2 (Müller and Strube, 2006) is used for multi-layer stand-off annotations whereas the Falko Excel Addin[6] is used for annotating target hypotheses. Automatic annotation relies on the UIMA[7] framework, which supports a modular integration of a wide range of NLP tools. For MERLIN, we have developed a specific SaltNPepper (Zipser et al., 2011) module for MMAX2 and custom converters for converting the transcribed data to PAULA, converting between PAULA and MMAX2, and processing the data with UIMA. For searching and visualizing the annotated corpus, the open source web-browser based search and visualization architecture ANNIS[8] is used.

The MERLIN corpus annotation is currently in progress. Strict interannotator agreement checks are being carried through. Level by level, 5% of the learner productions are annotated by all annotators. While a subset of them are discussed in the group and used to produce gold standard benchmarks, some productions are multiply annotated in a double-blind procedure in order to enable detailed analyses of inter-annotator agreement for the annotation of linguistic characteristics, target hypotheses, errors across multiple languages and proficiency levels.

6. Conclusion

In this paper, we presented the MERLIN corpus, a written learner corpus that has been designed to illustrate and validate the level system of the Common European Framework of Reference for Languages (CEFR) with authentic learner data. The MERLIN corpus provides learner metadata, detailed analytic and holistic CEFR ratings that have been thoroughly checked for reliability, and a comprehensive spectrum of learner and linguistic annotations along with full learner texts.

The selection of annotated features is solidly grounded in a user-needs study, SLA research, inductive learner text analyses, and an operationalization of CEFR scales. MERLIN users also have access to all tasks that were extracted from tests developed by ALTE-audited language testing institutions, which are each accompanied by a detailed description. A distinctive feature of MERLIN is that it covers German, Italian, and Czech, i.e., languages from three different linguistic families that, as an L2, have not yet received any consideration equivalent to English.

The complete annotated corpus will be freely available as a resource and through a didactically-oriented online platform (http://www.merlin-platform.eu) when the project concludes at the end of 2014.

7. References

Aharodnik, K., Chang, M., Feldman, A., and Hana, J. (2013). Automatic identification of learners' language background based on their writing in Czech. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1428–1436, Nagoya, Japan, October. Asian Federation of Natural Language Processing. http://aclweb.org/anthology/I13-1200.pdf.

Alderson, J. (1991). Bands and scores. In Alderson, J. and North, B., editors, *Language testing in the 1990s*, pages 71–86. British Council/Macmillan, London.

Alderson, C. (1996). Do corpora have a role in language assessment. *Using corpora for language research*, pages 248–259.

Bachman, L. F. and Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Bestgen, Y. and Granger, S. (2011). Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(2/3):235–252.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 17(1):47–64.

Bulté, B. and Housen, A. (2012). Defining and operationalising L2 complexity. In Housen, A., Kuiken, F., and Vedder, I., editors, *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, pages 21–46. Benjamins, Amsterdam.

[5] http://www.xmlmind.com/xmleditor/

[6] https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/marc/

[7] http://uima.apache.org

[8] http://www.sfb632.uni-potsdam.de/annis/

Carlsen, C. (2010). Discourse connectives across CEFR levels: A corpus-based study. In Bartning, I., Martin, M., and Vedder, I., editors, *Communicative Proficiency and Linguistic Development: intersections between SLA and language testing research*, pages 191–210.

Carlsen, C., editor. (2013). *Norsk Profil. Det europeiske rammeverket spesifisert for norsk. Et første steg*. Novus, Oslo.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., and Stede, M. (2008). A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues*, 49(2):271–293.

Council of Europe, (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. http;//www.coe.int/lang.

Eckes, T., (2009). *Many-Facet Rasch Measurement*. Strasbourg, France.

Fulcher, G., Davidson, F., and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1):5–29.

Fulcher, G. (2004). Deluded by artifices? The common european framework and harmonization. *Language Assessment Quarterly*, 1(4):253–266.

Gilquin, G. and Granger, S. (2010). How can data-driven learning be used in language teaching. *The Routledge handbook of corpus linguistics*, pages 359–370.

Gilquin, G., Granger, S., and Paquot, M. (2007). Learner corpora: The missing link in eap pedagogy. *Journal of English for Academic Purposes*, 6(4):319–335.

Glaboniat, M., Müller, M., and Wertenschlag, L. (2003). "Profile deutsch" - Lernzielbeschreibungen und sprachliche Mittel für Deutsch als Fremdsprache auf vier Niveaustufen des europäischen Referenzrahmens. In Schneider, G. and Clalüna, M., editors, *Mehr Sprache - mehrsprachig - mit Deutsch. Didaktische und politische Perspektiven*, pages 247–255. München: iudicium verlag 2003.

Granger, S. (2008). Learner corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus linguistics. An international handbook*, pages 259–275. Walter de Gruyter, Berlin, New York.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Aijmer, K., editor, *Corpora and Language Teaching*, pages 13–32. John Benjamins.

Hancke, J. and Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. In *Learner Corpus Research 2013. Book of Abstracts*, pages 54–56, Bergen, Norway. http://purl.org/dm/papers/Hancke.Meurers-13.html.

Hancke, J., Meurers, D., and Vajjala, S. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India. http://aclweb.org/anthology/C12-1065.pdf.

Hancke, J. (2013). Automatic prediction of CEFR proficiency levels based on linguistic features of learner language. Master's thesis, International Studies in Computational Linguistics. Seminar für Sprachwissenschaft, Universität Tübingen.

Hasil, J., Hájková, E., and Hasilová, H. (2007). *Brána jazyka českého otevřená*. Karolinum, Prague.

Hawkins, J. A. and Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge University Press.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, 91:663–667.

Jan, E., Schönherr, T., and Dallapiazza, R. M. (1998). *Tangram: Deutsch als Fremdsprache. Kurs- und Arbeitsbuch 1 A*. Hueber, Munich.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL Quarterly*, 45(1):36–62, March.

Lüdeling, A., Walter, M., Kroymann, E., and Adolphs, P. (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*, Birmingham.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In Walter, M. and Grommes, P., editors, *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweispracherwerbsforschung*, pages 119–140. Max Niemeyer Verlag, Tübingen.

Malvern, D. D., Richards, B. J., Chipere, N., and Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.

Meurers, D. (2013). Natural language processing and language learning. In Chapelle, C. A., editor, *Encyclopedia of Applied Linguistics*. Blackwell. http://purl.org/dm/papers/meurers-13.html.

Mezzadri, M. (2000). *Rete! Book 1*. Guerra Edizioni, Perugia.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. Peter Lang, Oxford.

Reznicek, M., Lüdeling, A., Krummes, C., and Schwantuschke, F., (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0*. http://purl.org/net/falko-maual.pdf.

Reznicek, M., Lüdeling, A., and Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In Díaz-Negrillo, A., Ballier, N., and Thompson, P., editors, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 101–123. John Benjamins, Amsterdam.

Roemer, U. (2008). Corpora and language teaching. In Lüdeling, A. and Kytö, M., editors, *Origin and history of corpus linguistics – corpus linguistics vis-à-vis other disciplines*, volume 1. De Gruyter Mouton.

Spinelli, B. and Parizzi, F. (2010). *Profilo della lingua italiana*. La Nuova Italia, Firenze.

Taylor, L. and Barker, F. (2008). Using corpora for language assessment. In *Encyclopedia of language and education*, pages 2377–2390. Springer.

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*, Atlanta, GA, USA, June. Association for Computational Linguistics.

Trosborg, A. (1995). *Interlanguage Requests and Apologies*. de Gruyter, Berlin.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163—-173. http://aclweb.org/anthology/W12-2019.pdf.

Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., Abel, A., and Hana, J. (2013). MERLIN: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings*, Florence, Italy. Libreriauniversitaria.it Edizioni.

Wisniewski, K. (2013). The empirical validity of the CEFR fluency scale: the A2 level description. In Galaczi, E. D. and Weir, C. J., editors, *Exploring Language Frameworks: Proceedings of the ALTE Krakow Conference*, Studies in Language Testing, pages 253–272. Cambridge University Press, Cambridge.

Wisniewski, K. (2014). *Die Validität der Skalen des Gemeinsamen europäischen Referenzrahmens für Sprachen. Eine empirische Untersuchung der Flüssigkeits- und Wortschatzskalen des GeRS am Beispiel des Italienischen und des Deutschen*. Number 33 in Language Testing and Evaluation Series. Peter Lang, Frankfurt.

Zipser, F., Zeldes, A., Ritz, J., Romary, L., and Leser, U. (2011). Pepper: Handling a multiverse of formats. In *33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft. DGfS-CL Poster Session*, Göttingen. http://purl.org/net/pepper-11.pdf.