International Conference "ICT for Language Learning" 6th edition

MERLIN: An Online Trilingual Learner Corpus Empirically Grounding the European Reference Levels in Authentic Learner Data

Katrin Wisniewski[1], Karin Schöne[1], Lionel Nicolas[2], Chiara Vettori[2], Adriane Boyd[3], Detmar Meurers[3], Andrea Abel[2], Jirka Hana[4]
[1]TU Dresden, [2]EURAC, [3]Univ. Tübingen, [4]Charles University Prague ([1,3]Germany, [2]Italy, [4]Czech Republic)
Katrin.Wisniewski@tu-dresden.de

Abstract

Since its publication in 2001, the Common European Framework of Reference for Languages (CEFR) has gained a leading role as an instrument of reference for language teaching and certification. Nonetheless, there is a growing concern about CEFR levels being insufficiently illustrated in terms of authentic learner data. Such concern grows even stronger when considering languages other than English (cf., e.g., Hulstijn 2007, North 2000). In this paper, we present the MERLIN project that addresses this need by illustrating and validating the CEFR levels for Czech, German, and Italian. To achieve its goal, we are developing a didactically motivated online platform to enable CEFR users to explore authentic written learner productions that have been related in a methodologically sophisticated and rigorous way to the CEFR levels. By making a significant number of learner productions freely accessible and easily searchable in a form that is richly annotated with linguistic characteristics and learner error types, the platform will assist teachers, learners, test developers, textbook authors, teacher trainers, and educational policy makers in developing a more comprehensive conceptualization of CEFR levels based on authentic learner data.

In the first, methodology-oriented part of this paper, we explain how the learner textual data were collected, re-rated, transcribed, double-checked and prepared for additional manual and automatic processing. We then illustrate the indicators we built to analyze L2 productions. Indicators were derived through (a) linguistic analyses of the performance samples, (b) the operationalization of the CEFR scale descriptors, (c) the study of relevant literature on SLA and language testing, (d) textbook analyses and (e) a questionnaire study. This study allowed us to devise a harmonized annotation schema taking into account both common and language-specific features (e.g., gender/article in German, reflexive possessive pronouns in Czech, pronoun particles in Italian).

In the second, application-oriented part, we explain how, by offering a large corpus of freely accessible empirical material, the project helps provide a fine-grained characterization of the CEFR levels and how it serves language teaching and learning. MERLIN thereby aims at responding to the suggestions of the Council of Europe itself, which solicits the development of supplementary tools for illustrating the CEFR levels (http://purl.org/net/CEFR-Goullier.doc). Furthermore, we explain how the platform enables the targeted users to retrieve authentic information about the relationship of the CEFR levels to a wide spectrum of well-defined, user-need-oriented L2 challenges. MERLIN users, such as teacher or learners, can thus compare their students' or their own performances and get a clearer picture of their strengths and weaknesses.

In the third, research-oriented part, we situate MERLIN with regards to two current topics in Second Language Acquisition: validation of CEFR scales and natural language processing for learner language.

[This publication reports on work from the MERLIN project, funded by the European Commission (518989-LLP-1-2011-DE-KA2-KA2MP). 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. Introduction

Though the Common European Framework of Reference for Languages (CEFR) nowadays holds leading role in language teaching and certification, it is often considered insufficiently illustrated in terms of authentic learner data (Fulcher 2004; Hulstijn 2007). The EU LLP project MERLIN (2012–2014), which brings together researchers from Germany, Italy, the Czech Republic, and Austria, is designed to improve this situation. Its major aim is 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 (www.merlin-platform.eu).

2. Methodology

2.1 Data selection and preparation

The MERLIN corpus was compiled using current best practices in data selection and preparation. The texts stem 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 and thus comply with international test quality standards. To compile the corpus, the written learner productions were extracted from the original tests and were re-rated, so as to guarantee a link as direct as possible to the CEFR scales, by trained raters according to a CEFR compliant analytical rating grid. This grid resembles Table 3 of the CEFR (CoE 2001: 29-30), adapted to the assessors' needs (Alderson 1991: 74). In addition, a holistic rating scale was used ('general linguistic range', Coe 2001: 110). Learners received, for each rating criterion, a direct CEFR level assignment resulting in competence profiles for all productions. The current corpus consists of roughly 200 texts per examination level (German: A1–C1, Italian A1–B2, Czech A2–B2). Test analyses, including Multi-Facet Rasch analyses, were carried out and showed a high degree of rating reliability.

2.2 Annotation workflow and technical background

When preparing an annotated text corpus, technical decisions can have far-reaching consequences as they impact how the corpus can be annotated and analyzed later. For MERLIN, the following decisions regarding transcription, format, manual and automatic annotation, and corpus exploration thus resulted from careful weighing of computational and explicit use-case considerations.

In order to transcribe the texts, we used Xml-mind, an XML-based editor. A dedicated style sheet was created with inline annotation related to the text structure and digitalization process. Once transcription was completed, all data was converted to PAULA (purl.org/net/paula), a standoff XML format designed as an exchange format for linguistic annotation. Further manual annotations are being performed with two tools: MMAX2 (mmax2.net) and the Falko Excel Addin (purl.org/net/falko). MMAX2 is a text annotation tool that allows multi-layered annotation and the Falko Addin is used for annotating target hypotheses (see sec. 2.3). Automatic annotation relies on the UIMA framework (uima.apache.org), which supports a modular integration of a wide range of NLP tools such as part-of-speech taggers and parsers. For searching and visualizing the annotated corpus, the open source web-browser based search and visualization architecture ANNIS (purl.org/net/annis) is used.

2.3 The MERLIN annotation scheme

The MERLIN texts are annotated with a wide range of language characteristics originating from various sources. First, there is a set of tags designed to determine whether the predictions of the CEFR scales correspond to learner behavior. There is insufficient research regarding this aspect of empirical validity (Fulcher 2004; Hulstijn 2007; Wisniewski –forthcoming-). Hence, selected CEFR scales were operationalized (Wisniewski 2012; -forthcoming-). Second, research-based annotation of coherence, grammar, vocabulary, orthography, and sociolinguistics (cf., e.g., Carlsen 2010, Bulté/Housen 2012, Lu 2011, Malvern et al. 2008, Bestgen/Granger 2011, Trosborg 1995) is being carried out. Third, in a questionnaire study and expert interviews, teachers and other envisaged user groups indicated specific CEFR illustration needs, so that the MERLIN Annotation Scheme identifies those properties, such as verbal aspect (Italian/Czech) or apostrophe use in (German/Italian). Fourth, textbook and language test analyses delivered further tags (e.g., German modal verbs). Finally, inductive analyses were carried out to identify additional language features. The sources of all tags will be transparently documented for MERLIN users.

The view of learner language as an evolving, individual interlanguage system in its own right is regarded as an important aspect of the MERLIN project and is reflected in the annotation scheme's linguistic annotations, which is complemented by learner error annotation reflecting a perspective more at home in a foreign language teaching and learning context. Target hypotheses are provided for each learner production to explicitly record the forms on which the annotated interpretations are based (Lüdeling 2008). The MERLIN Annotation Scheme also integrates the widely used 'target modification' dimension (cf. Díaz-Negrillo/Fernández-Domínguez 2006).

3. Applying MERLIN

3.1 MERLIN as illustration of CEFR levels

In order to be applicable across European languages, the descriptions of the CEFR levels needed to be general. However, it was recognized that the descriptors would 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, for example, by instigating the publication of the Reference Level Descriptions (RLDs) for national and regional languages (purl.org/net/rld). More and more RLDs tend to be based upon learner corpora, most prominently, the English (englishprofile.org), but also the Italian (Spinelli, Parizzi 2010) and the Norwegian Profiles (Carlsen 2013). While MERLIN similarly aims at illustrating CEFR levels, it differs by addressing three languages (supporting cross-language comparisons) and by providing access to the full texts, test tasks, and a wide range of linguistic and error annotations.

3.2 MERLIN for teaching, testing, and learning

The MERLIN user study revealed several use cases for the language classroom and beyond. First, 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. Teachers will be able to search the corpus for specific language-development milestones, e.g., the use of verb aspect or mood, but also identify typical obstacles to composing a special type of text or performing particular speech acts like requests. With the help of the platform, they can observe and trace those features on the different CEFR levels to better understand how learners advance in their learning process. Another scenario focuses on the use of standardized MERLIN ratings for revising personal or institutional assessment criteria and adjusting the rating behavior within a team of testers. To address the needs of the targeted users, the MERLIN platform interface will support selection and grouping of texts according to the proficiency level, the learners' L1, the type of text and the underlying task. 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.

3.3 MERLIN & research

3.3.1 MERLIN for the validation of the CEFR scales

Although the CEFR scales are used in an increasing number of contexts, even high-stakes ones, and despite many criticisms (e.g., Fulcher 2004; Hulstijn 2007), empirical validity has been little researched. The CEFR scales, originally calibrated with a sophisticated methodology, are nonetheless exclusively based on practitioners' beliefs about language ability (North 2000: 38). If rating scales claim empirical validity, though, they must reflect what learners actually do (Alderson 1991: 74). There is a well-known tendency of trained raters to base their decisions on scale-external aspects (Arras 2010; Eckes 2008). In order to examine empirical scale validity, it thus is not sufficient to look for empirical language features that are typical of *rated* CEFR levels. It is necessary to search for scale content correlates in learner language as directly as possible (Wisniewski 2012; -forthcoming-). By including operationalized CEFR level descriptions in its Annotation Scheme, MERLIN stands to contribute to the validation of the CEFR scales.

3.3.2 MERLIN to advance NLP of learner language

The MERLIN corpus provides valuable data for the development and evaluation of natural language processing tools for learner language (Meurers 2012). The corpus and its meta-information on learners and ratings readily support research on automatic native language identification, enabling such research to go beyond the current English learner focus. In a similar vein, the corpus has already been used for research on automatic proficiency classification for German (Hancke 2013). The MERLIN corpus also provides richly annotated learner data for the development and adaptation of NLP tools and applications that assist language learners in improving their vocabulary usage, coherence, spelling and grammatical accuracy.

4. Summary

In this paper, we presented the MERLIN project designed to provide a platform supporting the multi-faceted exploration of authentic written productions of learners of Czech, German, and Italian. We discussed the selection and preparation of the learner texts forming the empirical basis of the MERLIN corpus and the procedures used to obtain reliable CEFR ratings. We motivated the different

perspectives and needs which were integrated in the development of the MERLIN Annotation Scheme, and provided some background on the corpus representation and the manual and automatic annotation processes involved in preparing the corpus. The annotated MERLIN corpus will become freely accessible through a web-based platform at the conclusion of the project. It is designed to serve a range of practical and research purposes, from illustrating the CEFR levels for practitioners such as language teachers, curriculum designers or textbook writers to supporting the empirical validation of the CEFR scales and advancing the automatic NLP analysis of learner language.

References

[1] Alderson, J.C. (1991): Bands and scores. In: Alderson, J.C./North, B. (eds.): Language testing in the 1990s. London: British Council/Macmillan, 71-86.

[2] Arras, U. (2010): Subjektive Theorien als Faktor bei der Beurteilung fremdsprachlicher Kompetenzen. In: Berndt, A./Kleppin, K. (Hrsg.): Sprachlehrforschung: Theorie und Empirie – Festschrift für Rüdiger Grotjahn. Frankfurt: Lang, 169-179.

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

[4] Bulté, B./Housen, A. (2012): Defining and operationalising L2 complexity. In: Housen, A./Kuiken, F./Vedder, I. (eds.): Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. Amsterdam: Benjamins, 21-46.

[5] Carlsen, C. (2010): Discourse connectives across CEFR levels: A corpus-based study. In: Bartning, I./Martin, M./Vedder, I. (eds.): Communicative Proficiency and Linguistic Development: intersections between SLA and language testing research, 191-210. purl.org/net/Carlsen-10.pdf

[6] Carlsen, C. (Ed.) (2013): Norsk Profil. Det europeiske rammeverket spesifisert for norsk. Et første steg. Oslo: Novus.

[7] CoE 2001 = Trim, J./North, B./Coste, D.: Common European Framework of Reference for Languages: Learning, teaching, assessment. www.coe.int/lang

[8] Díaz-Negrillo, A./Fernández-Domínguez, J. (2006): Error-coding systems for learner corpora. In: RESLA 19, 83-102.

[9] Eckes, T. (2008): Rater types in writing performance assessments: A classification approach to rater variability. In: Language Testing 25 (2) 155-185.

[10] Fulcher, G. (2004): Deluded by Artifices? The Common European Framework and Harmonization. In: Language Assessment Quarterly 1 (4) 253-266.

[11] Fulcher, G./Davidson, F. (2007): Language Testing and Assessment. London/New York: Routledge.

[12] Hancke, J. (2013). Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language. Master's thesis, University of Tübingen.

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

[14] Lu, X. (2011): A corpus-based evaluation of syntactic complexity measures as indices of College-level ESL writers' language development. In: TESOL Quarterly 45 (1) 36-62.

[15] Lüdeling, A. (2008): Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Walter, M./Grommes, P. (Hrsg.): Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitsprachenerwerbsforschung. Tübingen: Niemeyer, 119-140.

[16] Malvern, D./Richards, B./Chipere, N./Durán, P. (2008): Lexical Diversity and Language Development. Quantification and Assessment. New York: Palgrave Macmillan.

[17] Meurers, D. (2012): Natural Language Processing and Language Learning. Encyclopedia of Applied Linguistics. Blackwell. purl.org/dm/papers/meurers-11.html

[18] North, B. (2000): The Development of a Common Framework Scale of Language Proficiency. Oxford: Peter Lang.

[19] Spinelli, B./Parizzi, F. (a cura di) (2010): Profilo della lingua italiana. Firenze: La Nuova Italia.

[20] Trosborg, A. (1995): Interlanguage Requests and Apologies. Berlin: de Gruyter.

[21] Wisniewski, K. (2012): The empirical validity of the CEFR fluency scale: the A2 level description. In: Galaczi, E.D./Weir, C.J. (eds.): Exploring Language Frameworks: Proceedings of the ALTE Krakow Conference. Cambridge: Cambridge University Press, 253-272. Studies in Language Testing.

[22] Wisniewski, K. -forthcoming-: 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. Frankfurt: Peter Lang. Language Testing and Evaluation Series.