# A TRILINGUAL LEARNER CORPUS ILLUSTRATING EUROPEAN REFERENCE LEVELS*

*Andrea* ABEL, *Katrin* WISNIEWSKI
*Lionel* NICOLAS, *Adriane* BOYD, *Jirka* HANA, *Detmar* MEURERS

**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 and for the development of curricula. Nonetheless, there is a growing concern about CEFR reference levels being insufficiently illustrated in terms of authentic learner data, leaving practitioners without comprehensive empirical characterizations of the relevant distinctions. This is particularly the case for languages other than English (cf. e.g. Hulstijn 2007, North 2000).
The MERLIN project addresses this demand to illustrate and validate the CEFR levels for Czech, German and Italian by developing a didactically motivated online platform that enables CEFR users to explore authentic written learner productions. The core of the multilingual online platform is a trilingual learner corpus composed of roughly 200 learner texts per CEFR level, produced in standardized language certifications validly related to the CEFR, covering the levels A1-C1.
The aim of this paper is to both present the MERLIN project with the motivation behind and its corpus and to discuss its current state.

**KEYWORDS** • Common European Framework of Reference for Languages (CEFR), learner corpus, learner language, language teaching, language certification

## 1. Introduction & Motivation

Though the Common European Framework of Reference for Languages (CEFR), above all its reference levels and descriptors, nowadays is perceived as *the* leading instrument for language teaching and certification, it is often considered insufficiently illustrated in terms of authentic learner data. Such concern grows even stronger when considering languages other than English (cf. e.g. Fulcher 2004, Hulstijn 2007, North 2000).

The EU LLP project MERLIN: "Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context" (2012-2014) is designed to improve this situation and to contribute to illustrating and validating the CEFR levels. It aims at researching and enhancing the empirical foundations of the CEFR scales by constructing a written learner corpus for selected languages, namely Czech, German and Italian as L2 (cf. Wisniewski/Schöne et al. 2013).

* This publication reports on work from the MERLIN project, funded by the European Commission (518989LLP12011DEKA2KA2MP). 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.

As the CEFR claims to be applicable across European languages, the descriptions of its levels had to be general. However, it was recognized that additional language-specific illustrations of the descriptors would be needed. In view of this demand to complement the CEFR, since 2001, the Council of Europe itself has encouraged the development of supplementary tools which better exemplify the features of a single language. One step in this direction was to instigate the publication of the Reference Level Descriptions (RLDs) for national and regional languages[1]. The tendency is that more and more RLDs tend to be based upon learner corpora, such as the English[2], but also the Italian (Spinelli/Parizzi 2010) and the Norwegian Profiles (Carlsen 2013). While MERLIN similarly aims at illustrating CEFR levels for given languages, it differs by following, for the first time, a multilingual approach. Thus, it addresses three languages from different families (Slavic, Germanic and Romance) and supports cross-language comparisons. In addition, it is distinct from related initiatives by providing free access to the full texts, test tasks, and a wide range of linguistic and error annotations on a didactically motivated online platform[3].

MERLIN also stands to contribute to the validation of CEFR scales. Although they are used in more and more contexts, even high-stakes ones, and despite well-founded criticism (cf. e.g. Alderson/Figueras/Kuijper/Nold/Takala/Tardieu 2006; Bausch/Christ/Königs/Krumm 2003; Alderson 2007; Fulcher 2004; Hulstijn 2007; Hulstijn/Alderson/Schoonen 2010; Little 2007) comparatively little research has been conducted regarding the (empirical) validity of the scale system. Although the scales were calibrated following a sophisticated methodology (North 2000), they exclusively reflect practicioners' beliefs about second language competence levels and (some of its) categories. While the scale research team has always been well aware of this mono-dimensional and thus somewhat limited validation perspective (North 2000: 38), a problematic trend currently observable consists in overestimating the meaning and the validity of these scales (cf. Wisniewski 2013, Wisniewski forthcoming). This 'reification' (Fulcher/Davison 2007: 232), a process that is described as "the propensity to convert an abstract concept into a hard entity"(Gould 1996: 27) is a dangerous process: learners' lives might be directly and profoundly affected by decisions taken on the basis of instruments that can only claim partial validity.

During CEFR scale calibration, the essential validity perspectives of building on sound theoretical constructs and, central for the MERLIN project, of matching the scales to authentic learner language (cf. Alderson 1991: 74) have been neglected or altogether disregarded. Although the above-mentioned RLD initiatives indeed do empirically illustrate the CEFR scales and detect language characteristics typical of the single levels, they rely on human ratings. It is by no means clear, though, if and to what degrees these ratings actually reflect the scale contents, as there is a tendency of even trained raters to base their decisions on scale-external, idiosyncratic aspects (Arras 2010; Eckes 2008; Pollitt/Murray 1996; Vaughan 1991). Thus, the RLD approach cannot serve as a control of empirical scale validity.

To avoid this logical circularity, it is necessary to search for scale content correlates in learner language as directly as possible and to the exclusion of ratings. To that purpose, in the MERLIN project CEFR level descriptions are operationalised, and learner texts will be analysed with the help of cluster and discriminant analyses (Wisniewski, forthcoming, Wisniewski 2013). Thus, MERLIN can contribute to researching the empirical validity of selected CEFR scales.

In this paper we want to introduce the MERLIN project, based on true desiderata from research and language teaching and certification (section 1), by presenting above all its corpus

[1] http://www.coe.int/t/dg4/linguistic/dnr_en.asp
[2] http://englishprofile.org
[3] http://www.merlin-platform.eu

as well as by detailing three key aspects: data collection and preparation (section 2), the creation of the annotation schemata (section 3), the annotation workflow, and the technical background (section 4). Finally, we will draw some conclusions and provide an outlook picturing the next steps as well as possible applications.

## 2. Data collection and preparation

A sophisticated methodology was devised to build the MERLIN corpus, starting from the data selection and preparation. The source of MERLIN texts are standardised, CEFR-related tests of German, Italian and Czech from long established testing institutions: German and Italian data from the telc institute in Frankfurt a.M., and Czech data from UJOP, Charles University in Prague and from UNIcert tests from the TU Dresden. 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, free written productions were extracted from the original tests, distributed as evenly as possible to different CEFR levels according to the availability of sufficient data at the testing institutions. Thus, the text collection consists of roughly 200 tests (3-4 different tasks per language and level) per examination level covering the levels A1-C1, for German, A1-B2 for Italian and A2-B2 for Czech. Furthermore, metadata, such as L1, gender and age, are available. Table 1 shows the distribution of all texts.

| | | German | Italian | Czech | |
|---|---|---|---|---|---|
| Basic User | A1 | 206 | 207 | 0 | |
| | A2 | 209 | 202 | 111 | |
| Independent User | B1 | 210 | 201 | 143 | |
| | B2 | 204 | 201 | 188 | |
| Proficient User | C1 | 204 | 0 | 0 | |
| | C2 | 0 | 0 | 0 | |
| | **TOTAL (texts)** | **1033** | **811** | **442** | **2286** |

Table 1: MERLIN original texts per CEFR level and language

In order to guarantee a link as direct as possible to the CEFR scales and to achieve a high concordance between the ratings of the three languages, all texts were re-rated (10% of them double-rated). Professional raters were trained to use the CEFR-based MERLIN rating grid which was developed on the basis of the experiences of the previous KOLIPSI initiative (cf. Wisniewski/Abel 2012). This grid resembles Table 3 of the CEFR (CEFR 2001: 29-30). It contains the CEFR scales (CEFR 2001 - Chapter 5: The user/learner's competences) for grammatical accuracy, vocabulary range, vocabulary control, coherence and cohesion, orthographic control, and sociolinguistic appropriateness, adapted to assessors' needs (Alderson 1991: 74) and made available in the three languages in question. In addition to the analytical scales, a holistic rating scale was used ("general linguistic range", CEFR 2001: 110). For each rating criterion, learners received a direct CEFR level assignment, resulting in competence profiles for all productions. Test analyses, including Multi-Facet Rasch (MFR) analyses, were carried through and showed a generally satisfying degree of rating reliability. Spearman's rho was between .607 for orthographic control and .832 for sociolinguistic appropriateness (Czech), between .847 for sociolinguistic appropriateness and .959 for general linguistic range (German),

and between 615 for sociolinguistic appropriateness and .908 for general linguistic range[4]. The Rasch analyses revealed high internal consistency values (separation reliability of .94 (Czech), .97 (German), and .97 (Italian). While the two Czech raters' severity was almost levelled, there were some differences between German raters, and considerable differences between the two Italian raters (reliability of rater separation index .00 (Czech), .98 (German), 1.0 (Italian))[5], with acceptable infit mean-square values throughout (Bond/Fox 2007)6. Fair averages which separate rater severity from examinee proficiency (Eckes 2009) were calculated for each test.

The following table presents the distribution of the texts that fulfilled quality requirements according to the thus corrected CEFR level as they are contained in the MERLIN corpus along with the total number of words (table 2). In other words, it shows the level at which test takers performed as opposed to the original level of examination taken (see table 1). As expected, the distributions differ substantially: As MERLIN raters could not only assign the originally targeted, but all CEFR levels to each written production, it becomes clear that many learners outperformed the examination level, while others achieved a lower CEFR level.

| | | German | Italian | Czech | |
|---|---|---|---|---|---|
| Basic User | A1 | 57 | 29 | 1 | |
| | A2 | 297 | 378 | 189 | |
| Independent User | B1 | 331 | 394 | 165 | |
| | B2 | 293 | 2 | 81 | |
| Proficient User | C1 | 42 | 0 | 2 | |
| | C2 | 4 | 0 | 0 | |
| | **TOTAL (texts)** | **1024** | **803** | **438** | **2265** |

Table 2: MERLIN re-rated texts per CEFR level and language as contained in the MERLIN corpus

In order to prepare the data to be included into the learner corpus, the hand-written texts had to be transcribed in a reliable way. To this end relevant criteria were established and detailed transcription guidelines were prepared. The texts were scanned and then transcribed with some basic on-the-fly inline annotations (for technical details see section 4) being carried out to record e.g. insertions, deletions, emphases, unreadable and ambiguous elements, foreign words, lists, emoticons, symbols and images, but also paragraphs. All personal and place names were anonymized with language-specific substitutions, direct citations from the text prompts as well as greetings and closings for responses in letter format identified. Furthermore, the transcriptions were checked against the original scans in a second run in order to eliminate transcription errors (e.g. anonymisations), and, at the same time tokenization errors were minimized (e.g. word boundaries not copied correctly). Quality control is a crucial aspect

[4] The Spearman rank correlation coefficient considers the extent to which a relationship between two variables may be described as a monotonic function and is commonly used for ordinal data (cf. Bachman 2004). Rho assumes values between -1 and +1; values close to -1/+1 indicate very strong, values close to 0 very weak correlations.

[5] Here, the separation index indicates the percentage of observed variance that cannot be explained by measurement error, thus showing how much raters differ from each other, with 0 meaning identical severeness and 1 highly differing severeness (Bond & Fox, 2007, Eckes 2009).

[6] The mean-square infit value is a measure of the model fit used in MFR. Infit values range from 0 to infinity. An infit value of 1.0 represents the exact amount of variance in the data expected from the model. Values differing from 1 indicate unexpected ratings (Eckes 2009). Bond & Fox (2007) define mean square values between 0.6 and 1.4 appropriate for rating scale data; all mean-square infits in the rater severeness facet of the MERLIN MFR fall into this range.

throughout the whole corpus creation process. The more accurate transcriptions are the more reliable the results of the following automatic and manual annotation steps will be (cf. Glaznieks/Nicolas/Stemle/Abel/Lyding 2014).

Last but not least during the preparatory phases a user modelling was carried out with the aims to discover users' needs concerning the MERLIN online platform both on a content as well as on a technical level. Thus, a comprehensive two-step online questionnaire study and expert interviews in the three project languages were carried through (115 concerning content and 55 concerning technical level were completed, and 6 resp. 3 experts were interviewed). The focus of the non-technical questionnaire was on users' views concerning features of learner language considered crucial but also on users' needs concerning the illustration of CEFR levels (e.g. vagueness etc.). Some results of the user modelling study were exploited also as a valuable source for the creation of the MERLIN annotation schemata (see section 3).

## 3. Creation of the annotation schemata

### *3.1. Linguistic analyses*

The MERLIN texts are annotated with a wide range of language characteristics originating from various sources. It was intended to identify meaningful indicators to describe aspects of learner language, errors and other linguistic characteristics that are not deficit-oriented, for German, Italian and Czech as a basis for data annotation, data analysis and data query. 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[7].

Indicators from four sources were incorporated:

a) CEFR-derived indicators.

First of all, it was deemed of particular importance to find out whether the predictions of the CEFR scales correspond to empirical learner behavior as there is insufficient research regarding this aspect of empirical validity of the CEFR scales (see section 1). Hence, where possible, the CEFR scales mentioned above were brought into a measurable form. In this operationalization process, exceedingly vague, self-referential, or subjective terms in the level descriptions had to be excluded (e.g. "Can sustain relationships with native speakers without […] requiring them to behave differently than they would with a native speaker", sociolinguistic appropriateness, B2, CEFR 2001: 122, but also aspects that were clearly related to spoken language only were ignored (e.g. "Can…keep up group discussions […]", sociolinguistic appropriateness scale, B2, CEFR 2001: 122) (cf. Wisniewski 2013; Wisniewski 2014). If, however, a level description mentions "greetings", "content jumps", "intelligibility", "idiomatic expressions" or "phrases" as characteristics of specific CEFR levels, these are annotated in the MERLIN corpus even if these so-called "scale variables" might not play a role in research or are often hard to clearly define. These annotations will allow to check the empirical relevance of the CEFR scales involved. It would be a sign of empirical validity if the scale contents were sufficiently salient and reliably observable in learner performances.

b) Research-based indicators.

Secondly, the MERLIN team tried to connect the learner productions at the differently rated CEFR levels to insights of Second Language Acquisition and Language testing research so

[7] The complete annotation scheme is available at http://www.merlin-platform.eu/

that many annotation tags are derived from an extensive literature review (e.g. Yang/Sun 2012, Vajjala & Meurers 2012, Paquot/Granger 2012, Bulté/Housen 2012, Bestgen/Granger 2011, Lu 2010, 2011, Mellor 2011, Carlsen 2010, Housen & Kuiken 2009, Bardovi-Harlig 2009, Malvern et al. 2008, Rimrott & Heift 2008, Burger 2007, Stede 2007, Nesselhauf 2005, Read/Nation 2004, Schmitt/Carter 2004, Dewaele 2004, Daller/Van Hout/Treffers-Daller 2003, Ortega 2003, Wray 2002, Bachmann 2002, Nation 2001, 2007, Read 2000, Wolfe-Quintero, Inagaki & Kim 1998, Laufer/Nation 1995, O'Loughlin 1995, Trosborg 1005, Halliday/Hasan 1989, Arnaud 1984). Research-based indicators include e.g.

- orthography: grapheme based errors, punctuation, capitalisation, 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, ...

Some of the identified indicators can also be used for the calculation e.g. of coherence/cohesion measures such as connector variety and accuracy, complexity and accuracy measures (cf. e.g. Bulté/Housen 2012, Wolfe-Quinteiro et al., Lu 2010, 2011) or vocabulary measures such as (Advanced) Guiraud's Index, theoretical vocabulary, Lexical Density or the percentage of error-free clauses. They take into consideration different dimensions of lexical knowledge, such as its variation, its sophistication, and its accuracy (Arnaud 1984, Daller/Van Hout/Treffers-Daller 2003, Laufer/Nation 1995, Malvern et al. 2008, Nation 2001, 2007, O'Loughlin 1995, Read 2000).

c) Experientially derived indicators.

A third source of indicators is represented by the envisaged future users of the platform like teachers and testers who in a questionnaire study and in expert interviews (see section 1) indicated specific CEFR illustration needs for the MERLIN annotation scheme to cover. Examples for such properties are verbal aspect for Italian and Czech, apostrophe use for Italian and German or the incorrect use of prepositions in general.

In addition, textbook analyses, such as *Tangram* for German (Dallapiazza 1998), *Rete!* for Italian (Mezzadri 2000) and *Brána jazyka českého otevřená* for Czech (Hasil 2007), revealed further crucial aspects of learning the three target languages. For example, German modal verbs or double negation in Czech were frequent topics in textbooks so that this is also a reason to include these aspects into the MERLIN annotation scheme and to make them explicitly searchable on the MERLIN platform.

d) Inductively derived indicators.

Lastly, linguistic analyses of performance were carried through in order to detect empirically relevant learner language features. In this step 10 texts per level and language were examined and relevant aspects revealed regarding e.g. 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.

Thus, a collection of a noticeable variety of indicators for German, Italian and Czech was created. The resulting harmonized annotation scheme contains a selection of meaningful indicators and is intended to strike the right balance between informativity and manageability

(cf. e.g. Díaz-Negrillo/Fernández-Domínguez 2006). The scheme takes into consideration both language-overarching features (e.g. register – level of formality; opening/closing formulae; collocations, idioms; grapheme errors), and some language-specific characteristics (e.g. gender/articles or modal particles in German, reflexive pronouns in Czech, pronoun particles and lexicalised clitics in Italian).

As already mentioned, the annotation scheme is not merely based on deficit-oriented error coding, but takes into account also other linguistic characteristics. Usual dimensions in error classification are a linguistic category classification (e.g. orthography, grammar, lexis …), a target modification classification (e.g. omission, addition, order …) and an error explanation classification (e.g. interference …) (cf. Lüdeling et al. 2005, Granger 2008, Granger 2002, Díaz-Negrillo/Fernández-Domínguez 2006, target modification dimension going back to Corder (1993) [1973]; further annotation dimensions are e.g. automatic annotations such as POS-annotation, see section 4, and error corrections using explicit target hypotheses, see section 3.2). Often, combinations are used, the linguistic classification being the most frequent one and hierarchies are adopted allowing e.g. for different linguistic levels and subcategories (cf. Díaz-Negrillo/Fernández-Domínguez 2006).

The annotations in MERLIN include a combination of the target modification dimension and a hierarchically structured linguistic classification of learner language features. For the latter, first of all a mandatory linguistic area has to be defined (specifically, along these categories: orthography, grammar, vocabulary, coherence/cohesion, sociolinguistic appropriateness, pragmatics, or a general category allowing to capture text and sentence intelligibility). The linguistic phenomenon is then further described on a second, compulsory annotation level (e.g., for the grammar: 'word order'), and, in some cases, also on an even more fine-grained third level (following the error-based example, the 'grammar'(level 1)/'word order' (level 2) definition must either get the tag 'word order in main clause' or the tag 'word order in subordinate clause' on level 3).

As to granularity, i.e. the number of tags or the aspects covered by single tags, existing error tagsets differ greatly from each other depending on the related project aims. Thus, the tagsets analysed by Díaz-Negrillo/Fernández-Domínguez (2006) reach from ca. 30 to 100 although due to different applications the number of tags cannot always be directly compared. The MERLIN annotation scheme contains 65 tags.

For many of these, the kind of target language modification can be annotated along the following six dimensions: choice, order, addition, omission, merge, split (as also in the German learner corpus FALKO[8], cf. Reznicek/Lüdeling et al. 2012). The tag names reflect the levels of annotation in order to improve clarity in the annotation process, e.g. C_Coh_jump includes the linguistic levels 1-3 (level 1: coherence/cohesion, level 2: coherence, level 3: content jumps) and e.g. G_Refl_pronrefl_O includes also the target modification level (level 1: grammar, level 2: reflexivity, level 3: reflexive pronoun, target modification: omission).

In addition to the tagset itself a detailed error tagging manual has been elaborated containing guidelines for annotators, namely descriptions of all the linguistic categories to be annotated, error coding principles and examples. This is important to enhance consistency in the annotation and minimize subjectivity as well as to increase inter- and intra-annotator agreement (cf. e.g. Granger 2003, Díaz-Negrillo/Fernández-Domínguez 2006). Furthermore, prototype examples (gold standard) have been elaborated with annotations to be used as training material, a training phase being a compulsory element for each MERLIN annotator.

[8] https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

The annotation scheme resp. the tagging manual records also further aspects such as the indicators source (which should also be made transparent for MERLIN end users), language specifity, need of target hypothesis 1 or 2 (see section 3.2), the tag span etc.

Finally, a comprehensive documentation has been established which describes the annotation scheme and the whole annotation procedure including measures for quality control (reliability checks) being important in each phase of the project[9].

### *3.2. Target hypotheses*

As every definition of aspects of learner language as erroneous must be hypothetical and often various interpretations are possible, and as, at the same time, all further analyses depend on these interpretations, it is regarded an aspect of quality assurance to make them explicit by formulating strictly rule-based, minimally intervening target hypotheses about what the learners most probably intended to say (cf. Lüdeling et al. 2008). According to Ellis (1994: 54) the target hypothesis is the necessarily assumed correct "reconstruction of those utterances in the target language". In the MERLIN project, target hypotheses are provided for each learner production to explicitly record the forms the annotated interpretations are based upon. Thus, all learner performances receive a new, additional layer which adheres to target-language rules. To that we widely follow the rules developed for the FALKO corpus adapting them to the project needs where necessary (cf. Reznicek/Lüdeling et al. 2012), and use the specifically developed FALKO Excel add-ins (see section 4). Thus, a target hypothesis I is formulated for each learner text targeting at minimal changes with the aim of constructing orthographically and grammatically acceptable, single sentences. This step is related to phenomena concerning linguistic correctness ("Systemfehler"), in contrast to phenomena related to the appropriateness of linguistic units ("Normfehler"), including e.g. stilistic variants (cf. Eisenberg 2007: 212, Schneider 2013: 30). Hence, the level of interpretation is expected to be comparatively limited and consistency between annotators as high as possible should be achieved by using detailed guidelines (cf. Reznicek/Lüdeling/Hirschmann in print, Reznicek/Lüdeling et al. 2012: 39). According to Reznicek/Lüdeling et al. (2012: 38) the provision of a target hypothesis I serves as normalisation level and is an important prerequisite for automatic processing steps. Furthermore, a target hypothesis II is elaborated with changes on a semantic and pragmatic level also considering of the context beyond the single sentence. Table 3 shows an example of a reconstructed sentence in Italian at target hypothesis I level.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ctok** | Gentili | Signori | , | Vi | scrivo | di | lamentare | a | causa | dei | vostri | servizi | brutti |
| **ZH1** | Gentili | signori | , | vi | scrivo | per | lamentarm | a | causa | dei | | servizi | brutti |
| **ZH1Diff** | | CHA | | CHA | | CHA | CHA | | | | DEL | | |
| **Engl. gloss (ctok)** | Dear | sirs | , | I write you to complain | | | | because of | | | your ugly services | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| che | abbiamo | ricevuto | in | vostro | villaggio | due | settimane | fa | . |
| che | abbiamo | ricevuto | nel | vostro | villaggio | due | settimane | fa | . |
| | | | CHA | | | | | | |
| which we have received | | | in your village | | | two weeks ago | | | . |

Table 3: MERLIN – Example of target hypothesis I for an Italian learner utterance (ctok = corrected token level, i.e. basically the transcribed learner text, ZH1 = target hypothesis I, ZH1Diff = differences concerning target language modification between ZH1 and ctok, automatically included by formulating ZH1)

[9] For reasons of space, in this paper we will not deal with quality control any further.

As FALKO has its focus on advanced learners of German and MERLIN, in contrast, contains many performances from beginners and intermediate learners (of German, Italian and Czech), by the way underrepresented in current learner corpus research as proficiency levels dominantly covered fall in the intermediate-advanced range (cf. Granger 2008: 264), some of the FALKO rules had to be adapted and new elements had to be added. An example of a newly added feature is the indication of the degree of certainty with which a target hypothesis could be formulated, an often very difficult task particularly on lower proficiency levels. Thus, three levels for marking arguable hypotheses are used (level 1: uncertain, but inferable; level 2: wild guess, level 3: no target hypothesis possible as partly not comprehensible; see table 4 for illustration).

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ctok** | Voi | -unreadable- | | su | anche | lei | poi | , | nel | futuro | . |
| **ZH1** | Voi | -unreada | assumerete | | anche | lei | poi | | in | futuro | . |
| **ZH1Diff** | | | INS | DEL | | | | DEL | CHA | | |
| **ZH1spec** | | | 2 | | | | | | | | |
| **Engl. Gloss (ctok)** | You | -unreadable- | | about | also | she | then, | , | in the | future | . |

Table 4: MERLIN – example for the indication of the degree of uncertainty (ZH1spec) at target hypothesis I level

The following table (table 5) provides an overview of the structure of the MERLIN annotation scheme:

| | | |
|---|---|---|
| **Tok** | = | automatically tokenized and manually checked learner text |
| **Correctly represented learner production** | | |
| Ctok | = | level for emergency corrections of tokenization/transcription |
| **Perspective Ia (orthography & grammar errors)** | | |
| TH1(ZH1) | | Target Hypothesis 1 (complete, corrected learner text) |
| ZH1Diff | = | level for marking differences between TH1 and ctok |
| ZH1spec | = | level 1-3 for marking speculative hypotheses |
| EA1_lev1,2,3 | | Error annotation 1, specified on level 1-3 |
| EA1_tlm | | target language modification |
| **Perspective Ib (vocabulary, coherence, sociolinguistic, pragmatic errors)** | | |
| TH2(ZH2) | | Target Hypothesis 2 |
| EA2_lev1,2,3 | | Error annotation 2, specified on level 1-3 |
| EA2_tlm | | target language modification |
| **Perspective II (learner language features that are not related to errors)** | | |
| LLF | | Non-error related learner language features |
| LLF_lev1-3 | | Specification of not error-related phenomena according to Annotation Scheme |

Table 5: Structure of annotations in MERLIN (version 12/2013)

While all MERLIN texts were annotated with target hypotheses I (TH1) and orthographic as well as grammatical errors (error annotation 1 – EA1), extended target hypotheses (TH2) and annotations of lexical, coherence, sociolinguistic, and pragmatic phenomena (EA2 and LLF) are available for an explorative core corpus.

## 4. Annotation workflow and technical background

Decisions such as the choice of software tools, data formats, and annotation procedures, may have substantial implications as they have important and long-lasting implications in terms of the broad and sustained usefulness of a corpus. For MERLIN, decisions regarding transcription, format, manual and automatic annotation, and corpus exploration resulted from careful weighing of computational and explicit use-case considerations.

The annotation workflow has been established with the following design criteria in mind:

- the building procedure was required to be reproducible and dynamically adaptable.
- the built corpus had to be extensible, of high quality and searchable.

### *4.1. Abstract workflow*

The abstract workflow (see figure 1) is organized in an iterative, user-oriented, fashion. It comprises three phases and five components that can all rely on one or several tools. The first phase is called the acquisition phase and comprises the component covering the process of obtaining a digital representation of the data. The second phase, the annotation phase, both includes the components that address the manual annotation tasks and the automatic ones with the help of human language technology (HLT) tools. The third phase, the exploration phase, comprises the component for corpus exploration that enables the linguist to explore the corpus and search for specific elements in context while the component for corpus statistics allows to compute general numerical values over the corpus (e.g. complexity measures). After the acquisition phase, the building procedure enters an iterative loop that sequentially alternates between the annotation and the exploration phases until both quantity and quality criteria are met.

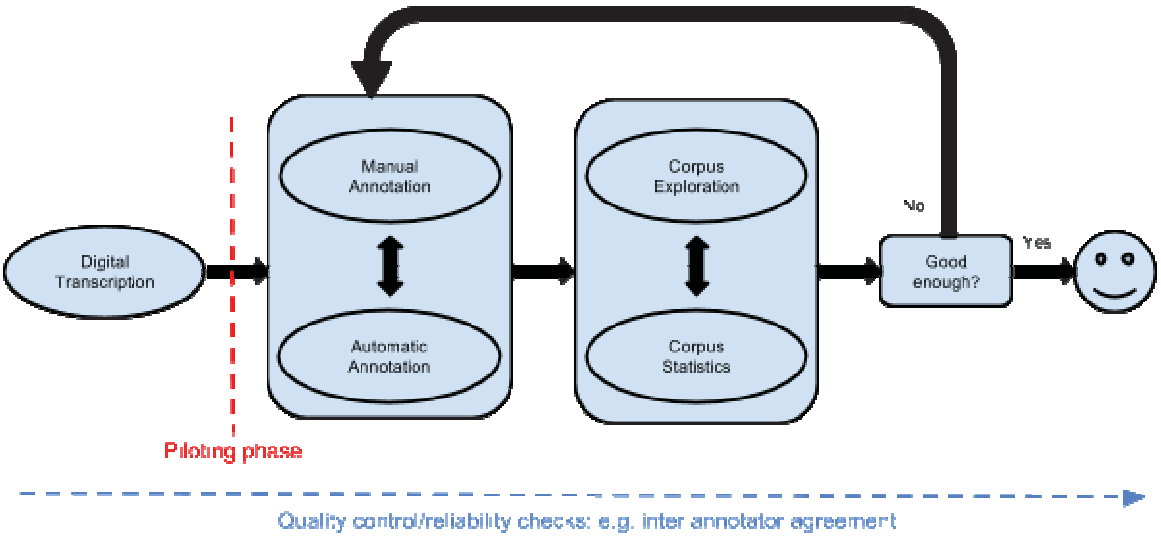


Figure 1: Abstract workflow

### *4.2. Implementation of the workflow*

In order to transcribe the texts, an XML-based editor, named Xml-mind[10], was used (see section 2). A dedicated style sheet was created in order to perform inline annotations as the text was being transcribed. The annotations performed related to the text structure and digitalization process.

Further manual annotations are performed with the help of two tools: MMAX2 (Müller/Strube 2006), a text annotation tool that allows multi-layered annotation, and the Falko Excel Addin [2] for annotating target hypotheses (see section 3). For automatic annotation, we rely on UIMA[11], a framework which supports a modular integration of a wide range of NLP tools such as part-of-speech taggers and parsers.

The open source web-browser based search and visualization architecture ANNIS (Zeldes/Ritz/Lüdeling et al. 2009) is used for implementing the corpus exploration component whereas regarding the corpus statistics component, we implement a range of different measures such as complexity measures (Hancke/Meurers/Vajjala 2012, Vajjala/Meurers 2012).

On a technical perspective, we decided to store the data, once the acquisition phase was completed, to the PAULA format (Christ 1994), a standoff XML format designed as an exchange format for linguistic annotation. For converting the data to and from the several format of the chosen tools, we intensively rely on the SaltNPepper convertion suite (Zipser/Romary et al. 2010) that already has conversion modules for PAULA, ANNIS and Falko. For the format of our UIMA tool chain, we created a custom converter, whereas for MMAX2, we chose to develop a new open-source SaltNPepper module.

Finally, we have chosen to use SVN for tracking changes in both the tools used and the data produced.

### *4.3. Compliance with the design criteria*

Tracking changes in both the tools used and the data produced allows for reproducibility of the annotation procedure. The iterative loop between the annotation and the exploration phases allows to dynamically evaluate the corpus and adapt the annotation procedure according to quantitative or qualitative evaluations. The format chosen for data storage, along with the manual and automatic tools selected, allows us to extend the corpus with new annotations when we see it fits. Finally, the corpus can be searched and browsed in an user friendly way thanks to the corpus exploration component.

## 5. Conclusion

The structure and contents of the MERLIN online platform contribute considerably to the exploration of learner language. Among the characteristics that distinguish MERLIN from other corpus initiatives is, firstly, the fact that the platform provides authentic learner texts in three languages that have not yet received any attention comparable to English as an L2 (Czech, Italian, German). It offers free online access and open source licenses for tools and resources. Also, full texts and tasks are visible along with an analytic, CEFR-based profile for every text. Meta-information is made available as well. These data are searchable for a wide spectrum of L2-related phenomena, some of which users indicated as important, while others are research-

[10] http://www.xmlmind.com/ (jan. 2014)
[11] http://uima.apache.org, (jan. 2014)

based, CEFR-related, or result from inductive text analyses. Users can sort texts according to CEFR level, tasks, or linguistic features, or they can create word lists or have selected statistical measures calculated.

The project team aims to guarantee a consistently high quality of the corpus by strictly controlling the data used and the processes applied in the project, the most important of which have been described in this article. Thus, for example, the texts stem from high-stakes language tests developed by prestigious testing institutions that undergo strict ALTE auditing procedures. The direct matching of each text to CEFR levels according to several criteria (like grammatical accuracy or sociolinguistic appropriateness) is carried through by professional raters, the ratings of whom are severely checked for reliability with state-of-the-art statistical procedures like Multi-Faceted Rasch analysis. Correspondingly, all steps of corpus compilation, like digitisation, the development of the annotation scheme, and manual or automatic annotations, are carefully documented and controlled for reliability and validity.

MERLIN is based on research in learner corpus linguistics, NLP, SLA, and language testing. For example, crucial validity issues regarding the scales of the CEFR can be analysed with the help of the MERLIN platform (cf. Wisniewski forthcoming, Wisniewski 2013), and automatic analyses of learner language with the help of natural language processing is advanced. In that perspective, this valuable data is of special interest for the development and evaluation of natural language processing tools for learner language such as automatic native language identification, automatic proficiency classification or computer-assisted learning (Meurers 2012, Hancke 2013). It thus represents a valuable resource for research in several fields, but MERLIN is also made for practitioners in language teaching, learning, and testing. The platform is meant to enhance their understanding of learner language, with a particular focus on its relation to CEFR levels which are commonly believed to be insufficiently illustrated. The future user groups are involved actively in the development of the annotation scheme and the creation of the interactive platform interface. MERLIN makes it easier to relate textbooks, examinations, and curricula to the CEFR and delivers realistic orientation to what can be expected on these levels.

This means that regarding the CEFR, MERLIN pursues a twofold objective: On a fundamental level, it contributes to further researching some of its most critical validity aspects. At the same time, MERLIN acknowledges the fact that the vast majority of descriptions of learner language in educational standards, curricula, language tests, and courses make use of the CEFR, and here, without dispute, guidance is needed. Hence, on a more practical level, MERLIN is meant to help base CEFR-related norms and decision on rich empirical data.

## REFERENCES

[CEFR 2001] Council of Europe (2001), *The Common European framework of reference for languages: Learning, teaching, assessment*, Cambridge, Cambridge University Press.

Alderson J.C. (2007), *The CEFR and the need for more research*, in "The Modern Language Journal" 91: 658-662.

Alderson J.C. et al. (2006), *Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of the Dutch CEFR Construct Project*, in "Language Assessment Quarterly" 3(1): 3-30.

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

Arnaud P.J.L. (1984), *The lexical richness of L2 written productionos and the validity of vocabulary tests*, in T. Culhane, C. Klein-Braley, D.K. Stevenson (eds.), *Practice and Problems in Language*, University of Essex Occasional Papers. Colchester, University of Essex: 113-148.

Arras U. (2010), *Subjektive Theorien als Faktor bei der Beurteilung fremdsprachlicher Kompetenzen*, in A. Berndt, K. Kleppin, (eds.), *Sprachlehrforschung: Theorie und Empirie – Festschrift für Rüdiger Grotjahn*, Frankfurt, Lang: 169-179.

Bachman L.F. (2004), *Statistical analyses for language assessment*, Cambridge, CUP 2004.

Bachmann T. (2002), *Kohäsion und Kohärenz: Indikatoren für Schreibentwicklung: Zum Aufbau kohärenzstiftender Strukturen in instruktiven Texten von Kindern und Jugendlichen*, Innsbruck, Studienverlag.

Bausch K.-R. et al. (eds.) (2003), *Der Gemeinsame Europäosche Referenzrahmen für Sprachen in der Diskussion. Arbeitspapiere der 15. Frühjarskonferenz zur Erforschung des Fremdsprachenunterrichts*, Tübingen, Narr.

Bardovi-Harlig K. (2009), *Conventional Expressions as a Pragmalinguistic Resource: Recognition and Productions of Conventional Expressions in L2 Pragmatics*, in "Language Learning" 59 (4) : 755-795.

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

Bond T.G., Fox C.M. (2007), *Applying the Rasch model: Fundamental measurement in human sciences*, Mahwah (NJ), Lawrence Erlbaum.

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

Burger H. (2007), *Phraseologie. Eine Einführung am Beispiel des Deutschen*. (3. Aufl.), Berlin, Erich Schmidt Verlag.

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

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

Christ O. (1994), *A modular and flexible architecture for an integrated corpus query system*, in *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, Budapest: 23-32.

Corder S.P. (1993 [1973]), *Introducing Applied Linguistics*, Harmondsworth, Pelican.

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

Daller H., van Hou R., Treffers-Daller J. (2003), *Lexical richness in spontaneous speech of bilinguals*, in "Applied Linguistics" 24: 197-222.

Dewaele J.-M. (2004), *Indiviual differences in the use of colloquial vocabulary. The effects of sociobiographical and psychological factors*, in P. Bogaards, L. Laufer (eds.), *Vocabulary in a second language*, Amsterdam, John Bejamins: 127-154.

Díaz-Negrillo A., Fernández-Domínguez J. (2006), *Error-coding systems for learner corpora*, in "RESLA" 19: 83-102.

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

Eckes T. (2009), *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Section H: *Many-Facet Rasch Measurement* (http://www.coe.int/t/dg4/linguistic/manuel1_en.asp, January 2014.)

Eisenberg P. (2007), *Sprachliches Wissen im Wörterbuch der Zweifelsfälle. Über die Rekonstruktion einer Gebrauchsnorm*, in "Aptum. Zeitschrift für Sprachkritik und Sprachkultur" 3/2007: 209-228.

Ellis R. (1994), *The study of Second Language Acquisition*, Oxford, Oxford University Press.

Fulcher G. (2004), *Deluded by Artifices? The Common European Framework and Harmonization*, in "Language Assessment Quarterly" 1 (4) : 253-266.

Fulcher G., Davidson F. (2007), *Language Testing and Assessment.* London/New York, Routledge.

Gould S.J. (1996), *The mismeasure of man*, London, Penguin.

Glaznieks A. et al. (2012), *Establishing a Standardised Procedure for Building Learner Corpora*, in "Apples – Journal of Applied Language Studies". Special Issue: Proceedings of LLLC2012.

Granger S. (2003), *Error-tagged learner corpora and CALL: a promising synergy*, in "CALICO Journal" 20 (3). Special issues on error analysis and error correction in computer-assisted language learning: 465-480.

Granger S. (2008), *Learner corpora*, in A. Lüdeling, M. Kytö (eds.), *Corpus linguistics: an international handbook* (Handbooks of linguistics and communication science; 29.1 _ 29.2), Berlin – New York: de Gruyter: 259-275.

Granger S. (2002), *A Bird's-eye view of learner corpus research*, in S. Granger, J. Hung, St. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam, John Benjamins: 3–33.

Halliday M.A.K., Hasan R. (1989), *Language, context and text: a social semiotic perspective*, Oxford, Oxford University Press.

Hancke J., Meurers D., Vajjala S. (2012), *Readability Classification for German using lexical, syntactic, and morphological features*, in *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*: 1063-1080.

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

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

Housen A., Kuiken F. (2009), *Complexity, Accuracy, and Fluency in Second Language Acquisition*, in "Applied Linguistics" 30 (4) : 461–473.

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.

Hulstijn J.H., Alderson C., Schoonen R. (2010), *Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them?*, in I. Bartning, M. Martin, I. Vedder (eds.), *Communicative Proficiency and Linguistic dvelopment: intersections between SLA and language testing research*, Eurosla Monograph Series (http://eurosla.org/monographs/EM01/EM01home.html)

Laufer B., Nation P. (1995), *Vocabulary size and use: lexical richness in L3 written production*, in "Applied Linguistics" 16: 307-322.

Little D. (2007), *The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Languages Education Policy*, in "The Modern Language Journal" 91: 645-655.

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.

Lu X. (2010*), Automatic analysis of syntactic complexity in second language writing*, in "International Journal of Corpus Linguistics" 15 (4) : 474–496.

Lüdeling A. (2008), *Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora*, in M. Walter, P. Grommes (eds.), *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitsprachenerwerbsforschung*, Tübingen, Niemeyer: 119-140.

Lüdeling A. et al. (2005), *Multi-level Error Annotation in Learner Corpora*, in S. Hunston, P. Danielsson (eds.), *Proceedings from the Corpus Linguistics Conference Series* (Corpus Linguistics 2005, Birmingham, 1415 July 2005) (http://www.corpus.bham.ac.uk/PCLC)

Malvern D. et al. (2008), *Lexical Diversity and Language Development. Quantification and Assessment*, New York, Palgrave Macmillan.

Mellor A. (2011), *Essay Length, Lexical Diversity and Automatic Essay Scoring*, in "Memoirs of the Osaka Institute of Technology", Series B Vol. 55, No. 2 (2011) : 1–14.

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

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

Müller Ch., Strube M. (2006), *Multi-Level Annotation of Linguistic Data with MMAX2*, in S. Braun, K. Kohn, J. Mukherjee (eds.), *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, Frankfurt, Peter Lang: 197–214.

Nation P. (2001), *Learning vocabulary in another language*, Cambridge, Cambridge University Press.

Nation P. (2007), *Fundamental issues in modelling and assessing vocabulary knowledge*, in H. Daller, J. Milton, J. Treffers-Daller, (eds.), *Modelling and Assessing Vocabulary Knowledge*, Cambridge, Cambridge University Press.

Nesselhauf N. (2005), *Collocations in a Learner Corpus*, Amsterdam, John Benjamins.

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

O'Loughin K. (1995), *Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test*, in "Language Testing" 12 (2) : 217-237.

Ortega L. (2003), *Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing*, in "Applied Linguistics" 24 (4) : 492–518.

Paquot M., Granger S. (2012), *Formulaic language in Learner Corpora*, in "Annual Review of Applied Linguistics" 32: 130-149.

Pollitt A., Murray N.L. (1996), *What raters really pay attention to*, in M. Milanovic, N. Saville (eds.), *Performance testing, cognition and assessment; Selected papers from the 15th Language Testing Research Colloquium*, Cambridge, Cambrudge University Press: 74-91.

Read J., Nation P. (2004), *Measurement of formulaic sequences*, in N. Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*, Amsterdam, John Benjamins: 23-35.

Read J. (2000), *Assessing vocabular*y, Cambridge, Cambridge University Press.

Reznicek M. et al. (2012), *Das Falko-Handbuch. Korpusaufbau und Annotation*en. Version 2.01. HU Berlin (http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01)

Reznicek M., Lüdeling A., Hirschmann H. (in print), *Competing Target Hypotheses in the Falko Corpus. A Flexible Multi-Layer Corpus Architecture*, in A. Díaz-Negrillo, N. Ballier, P. Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, Amsterdam, John Benjamins (Series Studies in Corpus Linguistics).

Rimrott A., Heift T. (2008), *Evaluating automatic detection of misspellings in German*, in "Language Learning & Technology" 11 (3) : 73-92.

Schmitt N., Carter N. (2004), *Formulaic sequences in action: An Introduction*, in N. Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*, Amsterdam, John Benjamins: 1-21.

Schneider J.G. (2013), *Sprachliche ‚Fehler' aus sprachwissenschaftlicher Sicht*, in "Sprachreport" 1-2/2013: 30-37.

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

Stede M. (2007), *Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik*, Tübingen, Narr.

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

Vajjala S., Meurers D. (2012), *On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition*, in J. Tetreault, J. Burstein, C. Leacock (eds.), *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7) at NAACL-HLT*, Montreal, Canada, Association for Computational Linguistics: 163–173.

Vaughan C. (1991), *Holistic assessment: What goes on in the rater's mind?*, in L. Hamp-Lyons (ed.), *Assessing Second Language Writing in Academic Contexts*, Norwood, Ablex: 111-125.

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

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*, Frankfurt, Peter Lang. Language Testing and Evaluation Series, 33.

Wisniewski K. et al. (2013), *MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data*, in *ICT for Language Learning, Conference Proceedings 2013*, Libreriauniversitaria.it Edizioni (http://conference.pixel-online.net/ICT4LL2013/common/download/Paper_pdf/322-CEF03-FP-Wisniewski-ICT2013.pdf)

Wisniewski K., Abel A. (2012), *Die Sprachkompetenzerhebung: Theorie, Methoden, Qualitätssicherung*, in A. Abel, C. Vettori, K. Wisniewski (eds.), *Gli studenti altoatesini e la seconda lingua: indagine*

*linguistica e psicosociale. / Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung*, I.1, Bolzano-Bozen, Eurac: 13-64 (http://www.eurac.edu/en/research/publications/PublicationDetails.aspx?pubId=0100156&type=Q)

Wolfe-Quintero K., Inagaki S., Kim H.-Y. (1998), *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu, Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Yang W., Sun Y. (2012), *The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels*, in "Linguistics and Education", 23 (1) : 31-48.

Wray A. (2002), *Formulaic Language and the Lexicon*, Cambridge, Cambridge University Press.

Zeldes A. et al. (2009), *Annis: A search tool for multi-layer annotated corpora. In Proceedings of Corpus Linguistics*, July 20-23, Liverpool (http://ucrel.lancs.ac.uk/publications/cl2009/)

Zipser F. et al. (2010), *A model oriented approach to the mapping of annotation formats using standards*, in *Workshop on Language Resource and Language Technology Standards*, LREC 2010.

**ANDREA ABEL** • Researcher and the coordinator of the Institute for Specialised Communication and Multilingualism, EURAC (Bolzano/Bozen), lecturer at the universities of Bolzano, Trento, Innsbruck and Regensburg. Research interests: modelling language competence and linguistic variation (written language, German as L1 and L2), research into learner and variety corpora, electronic and pedagogical lexicography, sociolinguistic aspects of language contact, languages and migration.

**E-MAIL** • andrea.abel@eurac.edu

**KATRIN WISNIEWSKI** • Researcher, Institute of Romance Studies, Technical University in Dresden. Coordinator of MERLIN. Research interests: modelling L2 competence, language testing (validity), learner corpus linguistics, spoken Language, German & Italian as L2, Applied Linguistics Methodology. PhD thesis on the validity of CEFR scales.

**E-MAIL** • katrin.wisniewski@tu-dresden.de

**DETMAR MEURERS** • Professor at Eberhard Karls Universität Tübingen, Germany and Universitetet i Tromsø, Norway. Research interests: Computational Linguistics, Second Language Acquisition, Linguistic Corpus Annotation, Learner Corpora, Linguistic Complexity and Readability Analysis, Native Language Identification, Lingustic Analysis of Syntax and Information Structure, Intelligent Language Tutoring Systems, Visual Input Enhancement.

**E-MAIL** • dm@sfs.uni-tuebingen.de

**JIRKA HANA** • Senior researcher at Charles University in Prague; research interests: computational processing of morphology, free word-order languages, clitics and learner corpora.

**E-MAIL** • jirka.hana@gmail.com