MERLIN
Multilingual Platform for the European Reference Levels: interlanguage exploration in context

User Manual
Version 1
11/2014

This project has been funded with support from the European Commission. This website reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Contents

Overview tables/graphics

PART I: Background information

1 The MERLIN project: aims and motivation................................................................. 3
2 Methodology.................................................................................................................. 5
   2.1 Data collection........................................................................................................ 5
   2.2 Transcription......................................................................................................... 5
   2.3 Re-ratings............................................................................................................. 6
   2.4 Manual annotation.............................................................................................. 6
      2.4.1 Target Hypotheses (TH)............................................................................... 7
      2.4.2 Annotation of learner language features..................................................... 9
   2.5 Automatized annotations..................................................................................... 15
   2.6 Quality control................................................................................................... 15

PART II: User guide

3 The MERLIN documentation section......................................................................... 18
   3.1 <<Using MERLIN>> ............................................................................................ 19
   3.2 <<MERLIN for research>> .................................................................................. 20
   3.3 <<MERLIN corpus>> .......................................................................................... 21
   3.4 <<MERLIN annotations>> .................................................................................. 25
   3.5 <<Documentation>> ......................................................................................... 27
   3.6 <<Download whole corpus>> ............................................................................. 29

4 The MERLIN search functionalities ........................................................................... 29
   4.1 <<Define a subcorpus>> ..................................................................................... 30
   4.2 <<Simple search>> ............................................................................................. 34
   4.3 <<Advanced search>> ......................................................................................... 36
   4.4 <<Statistics>> ..................................................................................................... 42

5 Help .......................................................................................................................... 46
   5.1 User manual....................................................................................................... 46
   5.2 Getting to know MERLIN: video introduction.................................................. 46
   5.3 MERLIN interface navigation: The screencast tutorial ...................................... 47
   5.4 How to apply MERLIN: usage scenarios.......................................................... 47
   5.5 Understanding the <<Advanced search>> output................................................. 47
   5.6 Glossary............................................................................................................. 49
   5.7 Frequently asked questions................................................................................. 49

References
Overview tables/graphics

<table>
<thead>
<tr>
<th>Part 1: Background information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Exemplary CEFR scale, “General linguistic range” (CoE 2001: 110)</td>
</tr>
<tr>
<td>Graphic 1</td>
<td>MERLIN annotation architecture</td>
</tr>
<tr>
<td>Table 2</td>
<td>Target hypothesis 1 (TH1), example</td>
</tr>
<tr>
<td>Table 3</td>
<td>Phenomena not captured on TH1, example</td>
</tr>
<tr>
<td>Table 4</td>
<td>Target hypothesis 2 (TH2), example</td>
</tr>
<tr>
<td>Table 5</td>
<td>Table 5: Annotated learner language features (tags &amp; definitions)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part 2: User guide</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphic 2</td>
<td>&lt;&lt;documentation&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 3</td>
<td>&lt;&lt;using MERLIN&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 4</td>
<td>&lt;&lt;using MERLIN&gt;&gt;, &lt;&lt;Using MERLIN for language teaching&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 5</td>
<td>&lt;&lt;MERLIN for research&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 6</td>
<td>&lt;&lt;MERLIN corpus&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 7</td>
<td>MERLIN test tasks for German (&lt;&lt;MERLIN corpus&gt;&gt;))</td>
</tr>
<tr>
<td>Graphic 8</td>
<td>Detail of a task description (&lt;&lt;MERLIN corpus&gt;&gt;))</td>
</tr>
<tr>
<td>Graphic 9</td>
<td>Number of texts taken on the different CEFR levels &amp; number of ratings on each CEFR level (&lt;&lt;MERLIN corpus&gt;&gt;))</td>
</tr>
<tr>
<td>Graphic 10</td>
<td>The MERLIN corpus in figures: Number of texts with target hypotheses 1/2 and error annotation 1/2 (&lt;&lt;MERLIN corpus&gt;&gt;))</td>
</tr>
<tr>
<td>Graphic 11</td>
<td>&lt;&lt; MERLIN annotations&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 12</td>
<td>MERLIN annotation structure, full corpus (&lt;&lt;MERLIN annotations&gt;&gt;))</td>
</tr>
<tr>
<td>Graphic 13</td>
<td>List of annotation tags with examples (&lt;&lt;MERLIN annotations&gt;&gt;))</td>
</tr>
<tr>
<td>Graphic 14</td>
<td>&lt;&lt;Documentation &gt;&gt;</td>
</tr>
<tr>
<td>Graphic 15</td>
<td>&lt;&lt;Download corpus&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 16</td>
<td>Search area</td>
</tr>
<tr>
<td>Graphic 17</td>
<td>Search interface &lt;&lt;Define a subcorpus&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 18</td>
<td>Specifying characteristics of a subcorpus, learner language features (&lt;&lt;Define a subcorpus&gt;&gt;))</td>
</tr>
<tr>
<td>Graphic 19</td>
<td>Exemplary output of &lt;&lt;Define a subcorpus&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 20</td>
<td>Exemplary output of &lt;&lt;view learner text and TH&gt;&gt; in &lt;&lt;Define a subcorpus&gt;&gt; output</td>
</tr>
<tr>
<td>Graphic 21</td>
<td>exemplary output of “view learner info and ratings” in &lt;&lt;Define a subcorpus&gt;&gt; output</td>
</tr>
<tr>
<td>Graphic 22</td>
<td>Search interface &lt;&lt;Simple search&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 23</td>
<td>Exemplary output of &lt;&lt;Simple search&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 24</td>
<td>Exemplary output of &lt;&lt;view learner info and ratings&gt;&gt; in &lt;&lt;Simple search&gt;&gt; output</td>
</tr>
<tr>
<td>Graphic 25</td>
<td>Search interface &lt;&lt;Advanced search&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 26</td>
<td>Defining POS tags, &lt;&lt;Advanced search&gt;&gt; (detail, German)</td>
</tr>
<tr>
<td>Graphic 27</td>
<td>Defining &lt;&lt;feature details&gt;&gt; in an &lt;&lt;Advanced search&gt;&gt; (detail, grammar)</td>
</tr>
<tr>
<td>Graphic 28</td>
<td>Exemplary lemma search, &lt;&lt;Advanced search&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 29</td>
<td>Exemplary lemma search output, &lt;&lt;Advanced search&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 30</td>
<td>Exemplary lemma search output, hidden meta information on left side&lt;&lt;Advanced search&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 31</td>
<td>Modify amount of context of a lemma in &lt;&lt;Advanced search&gt;&gt; output</td>
</tr>
<tr>
<td>Graphic 32</td>
<td>View automatic annotations in &lt;&lt;Advanced search&gt;&gt; output</td>
</tr>
<tr>
<td>Graphic 33</td>
<td>Search interface &lt;&lt;Statistics&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 34</td>
<td>Exemplary calculation of absolute frequency (all grammatical errors, German), &lt;&lt;Statistics&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 35</td>
<td>Output of exemplary calculation of absolute frequency (all grammatical errors, German), &lt;&lt;Statistics&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 36</td>
<td>All instances of exemplary frequency calculation (graphics 34-35) in &lt;&lt;Advanced search&gt;&gt; output</td>
</tr>
<tr>
<td>Graphic 37</td>
<td>Help</td>
</tr>
<tr>
<td>Graphic 38</td>
<td>Help in output of exemplary &lt;&lt;Advanced search&gt;&gt;</td>
</tr>
<tr>
<td>Graphic 39</td>
<td>&lt;&lt;about this search output&gt;&gt; help function in &lt;&lt;Advanced search&gt;&gt; output</td>
</tr>
<tr>
<td>Graphic 40</td>
<td>List of abbreviations used in the &lt;&lt;Advanced search&gt;&gt; output</td>
</tr>
</tbody>
</table>
Part I: Background information
1 The MERLIN project: aims and motivation

Introduction

The Common European Framework of Reference for Languages (CEFR) is the leading instrument for language teaching and certification in Europe. At its heart, although this is by no means its only contribution to standardization and vast improvements in language teaching and testing, there is the well-known system of CEFR levels, illustrated by exemplary scales. In spite of the very widespread use that is being made of these CEFR scales – there will hardly be a language test, a school curriculum, or a textbook without a reference to the scale levels – the scales are often 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, Wisniewski 2014).

The project MERLIN: “Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context” aims at improving this situation by offering a contribution to the illustration and the validation of the CEFR level system. MERLIN (2012-2014) was co-financed by the European Union (Lifelong Learning Programme, 518989-LLP-1-2011-1-DE-KA2-KA2MP). MERLIN aims at researching and enhancing the empirical foundations of the CEFR scales by constructing a written learner corpus for Czech, German and Italian as L2 (cf. Wisniewski et al. 2013, Abel et al. 2014).

Background: CEFR scales

The CEFR claims to be applicable across European languages. Thus, the level descriptions had to be general, like in the example below:

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.</td>
</tr>
<tr>
<td></td>
<td>Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events, but lexical limitations cause repetition and even difficulty with formulation at times.</td>
</tr>
<tr>
<td>A2</td>
<td>Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words.</td>
</tr>
<tr>
<td></td>
<td>Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information. Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulate about themselves and other people, what they do, places, possessions etc.</td>
</tr>
<tr>
<td></td>
<td>Has a limited repertoire of short memorised phrases covering predictable survival situations; frequent breakdowns and misunderstandings occur in non-routine situations.</td>
</tr>
<tr>
<td>A1</td>
<td>Has a very basic range of simple expressions about personal details and needs of a concrete type.</td>
</tr>
</tbody>
</table>

*Table 1: Exemplary CEFR scale, “General linguistic range” (CoE 2001: 110)*
For MERLIN, chapter 5 scales (‘communicative language competence’) were used (general linguistic range | grammatical accuracy | vocabulary range | vocabulary control| orthography | coherence & cohesion | sociolinguistic appropriateness). The CEFR is downloadable from the Council of Europe website.\footnote{http://www.coe.int/t/dg4/linguistic/cadre1_en.asp}

Illustration of CEFR levels

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 single languages. One step in this direction was to instigate the publication of the Reference Level Descriptions (RLDs) for national and regional languages. The tendency is that more and more RLDs tend to be based upon learner corpora, such as the English (www.englishprofile.org), 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. MERLIN also stands to contribute to the validation of CEFR scales.

Validation of CEFR level descriptions

The Council of Europe effort of scaling the CEFR descriptors (CoE 2001; North 2000; Schneider/North 2000) has led to immense improvements in standardization and transparency in language learning, teaching, and testing. Important decisions about language learners’ lives are taken with reference to the CEFR levels. One aspect that is yet insufficiently understood is the empirical validity of the CEFR scales (Fulcher 2004; Hulstijn 2007): If scales are used to describe or rate learner language, they must reflect what learners actually do (Alderson 1991).

As CEFR scales are increasingly used in high-stakes contexts, where important decisions about people’s lives depend on the interpretation of the CEFR scales (e.g., admission to University, naturalization), it is particularly important to be sure that the scales actually mirror empirical learner language. Here, very little research has been conducted (cf. e.g. Alderson et al. 2006; Alderson 2007; Fulcher 2004; Hulstijn 2007; Hulstijn et al. 2010; Little 2007; Wisniewski 2013, 2014).

CEFR scale calibration is based on practitioners’ beliefs about second language competence as expressed in ratings. However, it is not clear to what degree ratings actually reflect scale contents (Arras 2010; Eckes 2008; Pollitt/Murray 1996; Vaughan 1991). No learner language
analyses were carried through in the CEFR scaling process to support empirical validity. MERLIN aims at contributing to research regarding CEFR scale validity. Linguistic correlates to contents of central chapter 5 scales were operationalized and are searchable on the interface.

**Natural Language Processing (NLP) research**

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.

2 Methodology

2.1 Data collection

The MERLIN texts stem from the written production parts of CEFR-related, standardized high-quality tests from telc (Frankfurt/Main, Italian and German tests, www.telc.net) and ÚJOP (Prague, Czech tests, www.ujop.cuni.cz). These institutions are ALTE-audited (www.alte.org). The tasks were in use until 2013 and are now freely available on the platform. On this basis, a trilingual learner corpus was compiled that can be queried according to a variety of pre-determined aspects that were annotated manually and automatically (Glaznieks et al. 2014).

2.2 Transcription

The hand-written original learner texts were transcribed in an xml-based editor (xml mind©) by testing institutions (telc and ÚJOP). The transcribers followed transcription guidelines (available on the interface) and the reliability of the transcripts was checked, initially for a sample of 5% of the texts per CEFR level. As many transcription errors were detected, in the end almost all texts had to undergo a revision stage. The transcription guidelines included tags (inline annotation) for basic textual features such as unreadable or ambiguous stretches of language, foreign language words, emoticons, images, paragraphs, copied words from the rubrics, or greeting formulae. The anonymization (names, places) was part of the transcription process and was carried through based on the
guidelines. Transcripts served as a basis for annotations (see below). The transcription guidelines are available on the MERLIN interface (<<documentation>>, in German only).

2.3 Re-ratings

In the original tests the MERLIN texts were extracted from, test-takers received a score that was then weighted in different ways according to the importance attributed to writing in that particular test, leading to an overall pass or fail mark. For MERLIN, the procedure was necessarily different: the aim was to have a direct relation of texts to CEFR Chapter 5 scales of communicative language competence. Therefore, all texts were re-rated independently by professional raters.

The reliability of the re-ratings was examined with the help of Classical Test Theory and a Multi-Facet Rasch analysis. The latter is a probabilistic statistical procedure often used in language testing which allows for a correction of rating tendencies (e.g., leniency/harshness) and makes it possible to arrive at a fair average rating for each text. The intra-rater and inter-rater reliability was generally very high in MERLIN, with some exceptions for Italian. Therefore, the whole re-rating process was repeated for Italian resulting in a satisfying rating quality. The details can be found in the Technical Report (see <<documentation>> section).

In MERLIN, a holistic scale (based on the CEFR scale for general linguistic range) was used together with an analytical rating grid (rating criteria: orthography | grammatical accuracy | vocabulary range | vocabulary control | coherence & cohesion | sociolinguistic appropriateness) ranging from A1 to C2. Both instruments can be downloaded from the MERLIN <<documentation>> section. The fair average is calculated based on the holistic scale. If you compile your own corpus based on CEFR levels, these are also based on the fair average ratings. On the interface, you can access a rating profile with the original ratings for these rating criteria, as well.

Please note that many test-takers took a test which then turned out to be either too difficult or not much of a challenge to them. Therefore, on the MERLIN platform a distinction is made between the CEFR level of a test which need not be identical to the CEFR level(s) of the rating(s) (see also table 9 below. Both can be searched for separately (e.g., you can search for learners of German who took a B1 test but received only an A2 rating).

2.4 Manual annotations

Annotation is one of the core aspects of the MERLIN project. MERLIN has two types of annotations: ‘target hypotheses’ and annotations of learner language features. Where possible, automatized procedures (see 1.4) were used but most annotations were carried through manually.

The annotation was organized in 2 blocks:
2.4.1 Target Hypotheses (TH)

Research has shown that annotating learner language is a complex and partly speculative endeavor. Any mark-up of a learner language phenomenon requires a mental interpretation by the annotator. To guarantee transparency, coherence, and reliability of annotations, it is a good idea first explicitly write a ‘target hypothesis’ (TH), i.e. a corrected reconstruction of the learner text that a subsequent error annotation can build upon (Reznicek/Lüdeling et al. 2012). Also, target hypotheses are necessary for the successful implementation of many automated analyses (Díaz-Negrillo et al. 2010, Hirschmann et al. 2009). Thirdly, they can help future users of the MERLIN platform to understand annotations. MERLIN co-operates with the Falko project \(^2\)(Humboldt University, Berlin) which is one of the very few corpus initiatives that has a focus on target hypotheses and provides free access to the data. There are two types of TH1 (TH1 and TH2) in MERLIN which will be briefly explained in the following paragraphs.

**Target Hypothesis 1**

Target hypotheses for orthographic and grammatical errors (TH1) were written for the complete MERLIN corpus. In TH1 writing, the annotator is asked to change the learner text as little as possible in order to create a grammatically and orthographically correct version of the original learner text (‘minimal’ TH). In this table, you find an example:

\(^2\) https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko.
The following example by the same learner shows that in TH1, errors from other linguistic areas were ignored. There are content and technical reasons for this.

Table 2: Target hypothesis 1 (TH1), example

While the orthographical (capitalization error, word boundary error, missing hyphen) and grammatical (missing article) errors are corrected in the TH1, the lexically erroneous form "Reisespass" (instead of "Reisepass") was not substituted by another lexeme.

Target hypothesis 2

On a second level, extended target hypotheses (TH2) refer to aspects of sociolinguistic, lexical, and pragmatic deviations from what would normally be expected from a native speaker. TH2 thus aim at creating an acceptable version of the original learner text. For TH2, contextual aspects are taken into consideration. TH2 is an extension of TH1.

TH2 involve more subjectivity and difficulties in creating reliable decisions than TH1. This is a further reason to separate the two layers from each other. To illustrate the difference between TH1 and TH2, the following table might be useful:

Table 4: Target hypothesis 2 (TH2), example
2.4.2 Annotation of learner language features

An important guiding principle in MERLIN is the view of learner language as a system in its own that cannot be satisfactorily described with deficit-oriented error tags alone. It is important to stress that MERLIN annotation do include many error tags, but that particularly on EA2, there are annotation tags that record phenomena which are not errors, e.g. the realization of the speech act REQUEST or formulaic sequences.

The MERLIN annotation scheme thus represents a selection of meaningful, valid, and feasible features ('tags') that are manually annotated and that are supported by the MERLIN Computational Linguists team. It contains the following features:

### G_ Grammar

<table>
<thead>
<tr>
<th>G_Agr</th>
<th>agreement (subject and verb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G_Art</td>
<td>article</td>
</tr>
<tr>
<td>G_Clit</td>
<td>ITA: clitic</td>
</tr>
<tr>
<td>G_Conj</td>
<td>conjunction</td>
</tr>
<tr>
<td>G_Inflect_inexist</td>
<td>inexistdent inflection (nouns, adj, verb)</td>
</tr>
<tr>
<td>G_Morphol_wrong</td>
<td>wrong inflection (nouns, pronouns, adj)</td>
</tr>
<tr>
<td>G_Neg_negdoub</td>
<td>CZE: double negation</td>
</tr>
<tr>
<td>G_Neg_neggen</td>
<td>negation general</td>
</tr>
<tr>
<td>G_POS</td>
<td>part of speech error</td>
</tr>
<tr>
<td>G_Prep</td>
<td>preposition</td>
</tr>
<tr>
<td>G_Refl_pronrefl</td>
<td>reflexive pronoun</td>
</tr>
<tr>
<td>G_Refl_pronreplposs</td>
<td>CZE: possessive reflexive pronoun</td>
</tr>
<tr>
<td>G_Valency_complnumb</td>
<td>verb valency: number of obligatory arguments</td>
</tr>
<tr>
<td>G_Verb_asp</td>
<td>verb: aspect (CZE+ITA)</td>
</tr>
<tr>
<td>G_Verb_compl</td>
<td>verb formation (morphol.)</td>
</tr>
<tr>
<td>G_Verb_main</td>
<td>main verb</td>
</tr>
<tr>
<td>G_Verb_md</td>
<td>verb: mood</td>
</tr>
<tr>
<td>G_Verb_tns</td>
<td>verb: tense</td>
</tr>
<tr>
<td>G_Verb_vc</td>
<td>verb: voice</td>
</tr>
<tr>
<td>G_Wo_womaincl</td>
<td>word order in main clause</td>
</tr>
<tr>
<td>G_Wo_wosubcl</td>
<td>word order in subordinate clause</td>
</tr>
</tbody>
</table>

### O_ Orthography

<table>
<thead>
<tr>
<th>O_Abbrev</th>
<th>abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>O_Apostr</td>
<td>GER+ITA: apostrophe</td>
</tr>
<tr>
<td>O_Capit</td>
<td>capitalization</td>
</tr>
</tbody>
</table>
### O_Graph_act
CZE+ITA: diacritical marks

### O_Graph_graphgen
general grapheme error

### O_Graph_trans
grapheme transposition

### O_Punct
punctuation

### O_Wordbd
word boundary

### G_Intelligibility

<table>
<thead>
<tr>
<th>G_Intelltxt</th>
<th>intelligibility of text</th>
</tr>
</thead>
<tbody>
<tr>
<td>G_Intells</td>
<td>intelligibility of sentence</td>
</tr>
</tbody>
</table>

### V_Vocabulary

<table>
<thead>
<tr>
<th>V_FS</th>
<th>formulaic sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>V_Sequence</td>
<td>incomprehensible sequence caused by accumulation of lexical/grammatical error(s)</td>
</tr>
<tr>
<td>V_lexgrammerr_incompr</td>
<td>non-existing form (word or formulaic sequence)</td>
</tr>
<tr>
<td>V_form_nonexist</td>
<td>formulaic sequence: limited intelligibility</td>
</tr>
<tr>
<td>V_semdenot</td>
<td>semantic error: denotation (word or formulaic sequence)</td>
</tr>
<tr>
<td>V_semcon_att</td>
<td>semantic error: connotation (attitude), (word or formulaic sequence)</td>
</tr>
<tr>
<td>V_Word_semimprec</td>
<td>semantic error: precision (word or formulaic sequence)</td>
</tr>
<tr>
<td>V_Wordform_deriv</td>
<td>word formation error: derivation</td>
</tr>
<tr>
<td>V_Wordform_comp</td>
<td>word formation error: composition</td>
</tr>
<tr>
<td>V_FS_form</td>
<td>formulaic sequence: form error</td>
</tr>
</tbody>
</table>

### C_Coherence/Cohesion

<table>
<thead>
<tr>
<th>C_Con_accur</th>
<th>connector accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_Coh_jump</td>
<td>content jumps</td>
</tr>
<tr>
<td>C_Coh_ref</td>
<td>reference</td>
</tr>
<tr>
<td>C_Coh_txtstruct</td>
<td>metacommunicative device</td>
</tr>
</tbody>
</table>
S_ Sociolinguistic appropriateness

<table>
<thead>
<tr>
<th>Tag</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_Txt_grfw</td>
<td>salutations/complimentary closes</td>
</tr>
<tr>
<td>S_Txt_opcl</td>
<td>opening/closing formulae</td>
</tr>
<tr>
<td>S_Form_gen</td>
<td>inappropriate style (formality)</td>
</tr>
<tr>
<td>S_Form_addr</td>
<td>inappropriate addressing (formality)</td>
</tr>
<tr>
<td>S_Var_clit</td>
<td>ITA: lexicalised clitics (verbi procomplementari)</td>
</tr>
<tr>
<td>S_Var_duppron</td>
<td>ITA: personal pronoun redundancy</td>
</tr>
<tr>
<td>S_Var_synstr</td>
<td>ITA: marked syntactic structures</td>
</tr>
<tr>
<td>S_Var_che</td>
<td>ITA: 'che polivalente'</td>
</tr>
<tr>
<td>S_Var_woweil</td>
<td>GER: main clause word order after 'weil'</td>
</tr>
<tr>
<td>S_Var_partik</td>
<td>GER: modal particles</td>
</tr>
</tbody>
</table>

P_ Pragmatics

<table>
<thead>
<tr>
<th>Tag</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_Pol_dir</td>
<td>politeness - overly direct language form</td>
</tr>
<tr>
<td>P_Request_direct</td>
<td>direct REQUEST</td>
</tr>
<tr>
<td>P_Request_indirect</td>
<td>indirect REQUEST</td>
</tr>
</tbody>
</table>

Table 5: Annotated learner language features (tags & definitions)

These tags were chosen from a comprehensive list of features and indicators that were gathered in the project. Possible annotations were collected from:

1) the platform users’ perspective (based on a user study, on textbook and language test analyses)
2) the CEFR (by operationalizing elements of chapter 5 scales, CoE 2001)
3) Second Language Acquisition research (based on an extensive review of research literature)
4) learner texts (based on an inductive analysis of 10% of all learner texts)

There are many sources of information users of MERLIN can access with regard to the annotations:

- the bibliography that was used for choosing meaningful annotation tags in MERLIN, in the reference section of this manual
- the annotation scheme with all annotation tags that were implemented in MERLIN after a practicality check of a maximum list of annotations (available for download)
- documents like the MERLIN annotators’ manual (EN) and a documentation of additional annotation issues with fine-grained solutions to single annotation phenomena in the three project languages and other materials you can find on the interface (available on request).
- help functions on the interface, like a list of annotation tags with examples, screenshots explaining the annotation tiers, a glossary, and much more
- a black book that sums up experiences with annotations that future project in this field might want to work their way around (available for download)

**Annotation tags suggested by practitioners**

MERLIN is made for practitioners who work with the CEFR. Therefore, it is important to make the annotations as helpful for them as possible. To that end, a user study was carried through details of which are described in two reports available on the interface (<<documentation>>). The first part focused on the usefulness of content aspects, while the second part took into consideration the technical side of the MERLIN platform usability. The user study delivered important information on the annotation tags to be included. The annotation scheme contains information regarding the tags that stem from the user study.

A further possibility to understand users’ needs is the integration of aspects of L2 acquisition that are commonly treated in textbooks and in language tests into the MERLIN annotation scheme. In MERLIN, for example, "Tangram" for German (Dallapiazza 1998), "Rete!" for Italian (Mezzadri 2000) and "Brána jazyka českého otevřená" for Czech (Hasil 2007) were among the analyzed books. Also, the analysis of UJOP and telc language tests revealed certain notorious topics that delivered information for the annotation scheme.

Example annotations derived from these analyses include orthographical errors like erroneous capitalization, the incorrect use of the apostrophe in German and Italian, or grammatical errors such as the verbal aspect in Italian. In the lexical area, false friends or the use of idioms are recurrent topics that are mirrored in the MERLIN annotation scheme.

**Learner texts as source for deriving annotation tags**

10% of all learner texts in the corpus were analyzed by hand. This qualitative and inductive approach revealed a number of phenomena that were considered worthy of integration into the annotation scheme. Examples are word formation errors, many different types of problems related to the use of formulaic sequences, problems regarding the choice of the appropriate register in terms of formality or politeness. Interestingly, the majority of phenomena from this category belong to categories other than grammar and orthography.

**Research-based annotation**

A major effort went into an extensive research literature review for the different areas of language involved in MERLIN annotations. A full discussion of the single tags is not possible here, so we will try to give you a short overview. The references cited (and many more) can be found in the bibliography.
Annotations of grammatical phenomena include agreement, word order, negation, part-of-speech errors and much more. Thus, many accuracy and complexity measures can be calculated (see Lu 2010, 2011; Wolfe-Quintero et al. 1998, Ortega 2003, 2012, Houwen/Kuiken 2009). Orthography is an area that is not very extensively worked on in research. MERLIN annotation allows to access the relative/absolute frequency of different types of orthographic errors regarding, for example, punctuation, capitalization, or diacritics (Al-Jarf 2009, Bredel 2010, Cook 2005, Granger/Bestgen 2011, Perfetti/Rieben/Fayol 1997, Rimrott/Heift 2008, Sassoon 1995). These annotations are available for the whole corpus.

The MERLIN pilot core corpus annotations (EA2)

For a small subcorpus, aspects from other linguistic areas were annotated, too (EA2, error annotation 2). These subcorpus texts have obtained TH1 & EA1 and, in addition, TH2 and EA2. In the future, it would be desirable to have these explorative pilot annotations rechecked to then be able to cover the whole MERLIN database with TH2 & EA2 annotations. For now, we ask users to handle the core corpus annotations with caution.

Annotation tags for vocabulary take into consideration the manifold dimensions of lexical knowledge such as its accuracy, its depth, breadth, and sophistication (Nation 2001, 2007, Read 2000). Here, MERLIN has a strong focus on formulaic sequences (Wray 2002) which play a particular role in the acquisition of foreign languages (e.g., Pawley/Syder 1987, Schmitt et al. 2004). Many lexical tags regarding formulaic sequences are not error-related, but aim at capturing structures of particular interest in the acquisition process.

Sociolinguistic competence is defined as ‘the capacity to recognize and produce socially appropriate speech in context’ (Lyster 1994: 263). As for EA2 annotations in general, the annotation of phenomena pertaining to this competence (subjectivity, reliability) is methodologically challenging. The tags applied are in line with the project design and do not always comply with what is usually analyzed in the field (e.g., proficiency and amount of language contact, or diasystematic variation in learner texts as compared to L1 variation, Baker 2010, Baylea 2007, Bayley/Regan 2004, Biber/Finegan 1994, Hudson et al. 1005, Hymes 1974, Mougeaon/Dewaele 2004, Regan et al. 2009, Van Compernrolle/Williams 2012, Yu 2012, Zuskin 1992). In MERLIN, the appropriateness of language forms with regard to ‘formality’ is annotated (e.g., substandard forms, use of forms that pertain to oral communication, overly formal language, see Koch/Oestereicher 2011), and tags that are text-type tailored such as greetings or opening and closing formulae are integrated. Single language-specific variational aspects chosen reflect choices of structures that either do not pertain to the written language and/or seem to be on their way to be accepted as standard variants, but would not normally be accepted in the task types they can be found in in the MERLIN texts.


CEFR-based annotation

To find out if the CEFR scales reflect learner language, it is important to operationalize their descriptors without making use of human ratings which have often turned out to not be based on rating instruments even when they are reliable (Eckes 2008, Wisniewski 2010). If scale descriptors are put into a measurable form like this, the relationship between selected CEFR scales and learner language becomes much clearer.

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, CoE 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, CoE 2001: 122) (cf. Wisniewski 2013, 2014). If, however, a level description mentions “greetings”, “content jumps”, “intelligibility”, “idiomatic expressions” or “phrases” as characteristics of specific CEFR levels, these features were checked for feasibility in the MERLIN annotation scheme, even if these so-called “scale variables” might not play a role in research or are often hard to clearly define. These annotations 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. MERLIN cannot offer a complete validation of CEFR scales, but it focuses on a selection of meaningful aspects.

Accessing annotations on the interface

Annotations can be accessed directly via the search functions of the interface (<<Advanced search>>, <<Define a subcorpus>>), where every single occurrence is displayed in context.

Another possibility to access annotations is to use them for statistical measures. The simplest possibility is to count the total number of annotation tags occurring in a (sub)corpus. In many cases, though, it is more meaningful to use annotations for the calculation of normalized measures (i.e., in MERLIN, per sentence, T-unit, or token). To give you an example, this allows you to compare the average number of morphological errors per sentence in Czech B1 vs Czech B2 learner texts. Normalized measures of manually annotated phenomena are calculated on the basis of automatized segmentation procedures (see below). The frequency-based measures are available in the <<statistics>> section on the interface, along with a number of more complex measures of complexity.
2.5 Automatized annotations

The automatic annotation in the MERLIN corpus serves to support the manual annotation and to make accessible a wide range of linguistic features for the calculation of indicators and in direct corpus searches.

The automatic annotations for the MERLIN corpus can be divided into four categories:

1. Linguistic units needed for the manual annotation:
   a) tokens
   b) sentences

2. Linguistic units needed for the calculation of measures:
   a) t-units
   b) a range of clause types

3. Linguistic annotation using existing definitions and tools
   a) part-of-speech
   b) lemma
   c) constituency and dependency parses

4. Linguistic annotation with MERLIN-specific definitions and tools
   a) repetitions within texts
   b) citations of task material

We have applied existing automatic annotation tools developed for the target languages in order to expand the range of available linguistic annotation beyond what would have been possible with time-consuming and expensive manual annotation. However, it is important to keep in mind that automatic annotation is particularly challenging for learner language, since learner language often deviates considerably from the target language across all levels of linguistic analysis, from spelling to semantics.

The following tools were used for all three MERLIN languages:

Texts were tokenized using the tokenizer for Indo-European languages from LingPipe and the resulting tokenization was then corrected by hand. Sentences were annotated with the OpenNLP sentence segmenter. Repetitions were identified using the Saphre library on the basis of the automatic part-of-speech and lemma annotation.

Please refer to the <<MERLIN for research>> section to learn more about the language-specific tools used for automatic annotation.

2.6 Quality control

In order to organize and control annotation reliability, a number of measures were taken. All instruments (TH 1 & TH2 rules, annotation scheme for EA1 and EA2) were piloted and revised before their implementation. Piloting was organized in two steps. First, the
The annotation procedure itself was tested to get a first idea of how to concretize and change the annotation guidelines and instruments. Then, in a piloting process, a restricted number of texts were annotated by all annotators of a language team in order to again check the practicality and the sufficiency of the annotation guidelines and in order to detect possible technical problems.

Secondly, all annotations are based on guidelines (annotator manual, see interface). The guidelines are enriched by fine-grained decisions on single aspects of annotation (document on additional annotation issues, see interface). Thirdly, the reliability of the annotations is controlled. Reliability of annotations was controlled for 5% of the texts on each test level for target hypotheses and error annotation. Different methods were applied:

In a qualitative approach, half of the files are annotated independently by the coders to then be commonly discussed with the aim to arrive at a consensus. These texts served as a reference throughout the annotation process. The qualitative approach turned out to be extremely important for a common understanding of the annotation scheme. In a double-blind procedure, the second half of the files checked for reliability was annotated by all coders without their knowledge. The annotations in these files were checked for coder reliability qualitatively and quantitatively.
PART II: User guide
3. The MERLIN documentation section

The macrostructure of the freely accessible MERLIN platform is organized in a documentation area (vertically placed on the left of the interface, see graphic 2) and a search area (horizontal search bars, see graphic 23). Chapter 3 explains the most important contents and functions of the documentation section.
3.1 <<Using MERLIN>>

In the <<using MERLIN>> section, you get general advice on possibilities for applying MERLIN in different professional settings. This section is useful to give you a first idea of what the project is all about.

You can unfold the chapters (click ) to learn more about, for example, how you can use MERLIN for language teaching or for developing teaching materials.
3.2 <<MERLIN for research>>

The section <<MERLIN for research>> is meant for anyone interested in background information regarding different aspects of the project. Whenever you browse through the interface and would like to learn more about the project rationale, it is advisable to consult the <<MERLIN for research>> section.

![Graphic 5: <<MERLIN for research>>](image)

As graphic 5 shows, there is a wealth of information regarding the linking of MERLIN texts to the CEFR. All relevant documents like the rating grids, the tasks, the technical report regarding the quality of the ratings are accessible from here. Also, the workflow that the MERLIN data underwent is outlined in this section: you can find out how the transcription and the annotation were carried through, with the help of which tools and schemes. You get information on quality control aspects of the manual and automatic annotations as well.

Thirdly, there is information regarding the possibilities to use MERLIN in researching the validity of the CEFR scales, second language acquisition, and Natural Language Processing of learner language. You also find a list of relevant references.
The section `<<MERLIN corpus>>` yields much information on the MERLIN data. You can learn more about the texts and the testing institutions, you can download the rating grids that were used, and you can see a list of the test tasks:
**Test tasks**

We provide a comprehensive overview of the test tasks by target language and CEFR level tested. The level of the test may differ from the level that the learner text received in the re-ratings. The tasks are represented using a grid that was developed for these purposes by ALTE (Association of Language Testers in Europe, [www.alte.org](http://www.alte.org)). The grid contains detailed information about the tasks and the specific characteristics of the intended text, e.g., regarding topic, register, domain (author: Olaf Bärenfänger).

### General notes on task descriptions

**German**

**A1**
- Informal e-mail: ask a friend for help with finding an apartment
- Informal e-mail: arrange an appointment with a friend to go swimming together
- Informal letter: congratulate to birth of a child

**A2**
- Formal letter to housing office
- Informal letter: ask friend to take care of pet
- Informal letter: offer a ticket not used to a friend

**B1**
- Informal letter for New Year to a friend
- Informal letter to a friend announcing a visit
- Informal letter: birthday congratulations

**B2**
- Formal letter: ask for information at Au pair Agency
- Formal letter: Au pair writes letter of complaint to Agency
- Formal letter: apply for internship in sales department

**C1**
- Essay: why it’s of value to learn German
- Online article: about sticking to one’s traditions and “assimilation” in a new environment
- Report: about the housing situation

**Graphic 7: MERLIN test tasks for German (<<MERLIN corpus>>)**

If you click on a task a pdf will open in an extra window that contains the task itself along with a detailed task description that is based on a Grid developed by the Association of Language Testing in Europe (ALTE, [www.alte.org](http://www.alte.org)). The task description tells you more about the length of the task, the type of language in the expected response, its difficulty and much more.
The "MERLIN corpus" section also provides information on available metadata (like age, gender, or mother tongue), and there is a table with the total numbers of texts available per test level and per rated overall CEFR level:

**The MERLIN corpus in figures**

Number of texts per CEFR level of the test (test level) compared to the number of texts per CEFR level assigned in the re-rating (fair average)

<table>
<thead>
<tr>
<th>Language</th>
<th>Test Level</th>
<th>Fair Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>A1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>C1</td>
<td>2</td>
</tr>
<tr>
<td>Italian</td>
<td>A1</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>A1</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>209</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>210</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>C1</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>4</td>
</tr>
<tr>
<td>German</td>
<td>A1</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>376</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>394</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>A1</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>331</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>C1</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>2286</td>
<td>2285</td>
</tr>
</tbody>
</table>

Graphic 9: Number of texts taken on the different CEFR levels & number of ratings on each CEFR level ("MERLIN corpus")
And a table with information regarding the depth of annotation of the MERLIN texts:

<table>
<thead>
<tr>
<th></th>
<th>Czech</th>
<th>German</th>
<th>Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texts</td>
<td>442</td>
<td>1033</td>
<td>813</td>
</tr>
<tr>
<td>TH1</td>
<td>440</td>
<td>1033</td>
<td>813</td>
</tr>
<tr>
<td>EA1</td>
<td>361</td>
<td>752</td>
<td>754</td>
</tr>
<tr>
<td>TH2</td>
<td>231</td>
<td>275</td>
<td>154</td>
</tr>
<tr>
<td>EA2</td>
<td>198</td>
<td>258</td>
<td>85</td>
</tr>
</tbody>
</table>

*Graphic 10: The MERLIN corpus in figures: Number of texts with target hypotheses 1 & 2 and error annotation 1 & 2*
This section contains information regarding MERLIN annotations, with a focus on the manual annotations.

You can get an overview of the annotation architecture:

The MERLIN data have been enriched with a multi-level annotation. NLP (Natural Language Processing) was used for automatic learner language annotations such as tokenization and lemmatization, part-of-speech tagging or segmentation into sentences or T-units.

Annotations in the full MERLIN corpus

The main annotations available for the whole MERLIN corpus are target hypotheses (target hypotheses 1) and annotations of grammatical and orthographical learner language features (error annotation 1).
Also, you get access to the **annotation scheme** all manual MERLIN annotations are based upon, with examples for all three languages, a glossary for less common terminology, and clear definitions of each tag. Furthermore, you can download a comprehensive **progress documentation** of difficult questions that arose regarding single annotation aspects. If you come across an annotation that seems questionable to you, this document might be useful (it is also accessible from <<documentation>>).

If you do not want to read through the full tag definitions, instead of downloading the complete annotation scheme, you can also consult a list with all annotation tags and examples:

**List of learner language features with examples**

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Orthography</th>
<th>Intelligibility</th>
<th>Vocabulary</th>
<th>Cohesion/Coherence</th>
<th>Sociolinguistic appropriateness</th>
<th>Pragmatics</th>
</tr>
</thead>
</table>

**GRAMMAR TAGS**

<table>
<thead>
<tr>
<th>word order in main clause</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>“[Vielleicht du könntest mir bei meiner Wohnungssuche helfen.]”</td>
<td>*[Solist du Wasser und Eikorn mitbringen.]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>word order in subordinate clause</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>“[Wenn haben Sie Zeit, dann bitte sagen Sie mir.”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>negation general</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Ich habe [nicht] Zeit.”</td>
<td>“Er wird dort arbeiten [nein].”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CZE: double negation</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>“[mám] žádný čas (nemám žádný čas); [nikdo] [volal] [nikdo nevolal]”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>verb valency: number of obligatory arguments</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZE: “Pět vstává v 6 hodin. On nezná, protože on nemá hlad.”</td>
<td></td>
</tr>
<tr>
<td>GER: “Er hat uns nicht gesagt, ob (er) kommen will.”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>agreement (subject and verb)</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Jana [hast] gelesen,”</td>
<td>“Jana [sind] müde”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>reflexive pronoun</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZE: “[síma] [sí],”</td>
<td>“Er [entschuldigt], “Laura und Ferdinand reden [sich]”</td>
</tr>
<tr>
<td>GER: “[sí] [sí] eine gute Mutter”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CZE: possessive reflexive pronoun</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>“[poříbuj] [moji] knihu,”</td>
<td>“[vidím] [máho] otce”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>inexistet inflection (nouns, adj, verb)</th>
<th>Example*</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjective: “[ein] [blauer] Himmel”</td>
<td>“[feuer] [feuer]”</td>
</tr>
<tr>
<td>noun: “[groß] [großer]”</td>
<td>“[fahrrad] [fahrrad]”</td>
</tr>
</tbody>
</table>

**Graphic 13: List of annotation tags with examples (<<MERLIN annotations>>)**
Here, all freely available MERLIN-related documents can be found, i.e.:

- reports produced in the project, e.g. on the quality of the ratings and the usability of the platform (user studies)
- all test tasks used
- grids used for rating;
- guidelines and schemes used for transcription and annotation (e.g., annotation scheme)
- publications, presentations and the like by the MERLIN team
...
This section gives you an immediate download option for the whole MERLIN corpus. You can download zipped files for Czech, Italian, and German in .txt or .pdf. You can decide whether you want to download the original learner texts with metadata (e.g., L1, age) or whether you also want to include target hypotheses.

If you are interested in a more specific collection of texts, go to <<define a subcorpus>> first. There, you can create your customized subcorpus, download it, search in it (<<Simple/Advanced search>>) or have statistical measures displayed (<<statistics>>).
4. The MERLIN search functionalities

The search options include:
- a simple search
- an advanced search
- a “define a subcorpus” section
- a statistics section
4.1 <<Define a subcorpus>>

It is recommendable to start using the search option by defining a collection of texts one is interested in (<<Define a subcorpus>>, see graphic 17):

Here, it is possible to specify criteria according to which a collection of learner texts (a so-called subcorpus) can be tailored to users’ needs. The criteria are:

- target language of the text
- original CEFR level of the test
- fair average CEFR rating of the texts (see section 2)
- one or more CEFR levels of single rating criteria (grammatical accuracy | vocabulary range | vocabulary control | sociolinguistic appropriateness | coherence & cohesion | orthography)
- test task

Furthermore, it is possible to sort text according to learner information:

- mother tongue (L1)
- age
- gender
Also, texts can be selected according to up to three learner language features and/or words:

![Filter for words and learner language features](image)

**Graphic 18: Specifying characteristics of a subcorpus, learner language features(<<Define a subcorpus>>)***

The subcorpus needs to be given a name that can be chosen by you; by clicking on “define subcorpus and show texts”, this subcorpus will be available for further searches for 24 hours. After that, you will have to redefine the subcorpus.

**<<Define a subcorpus>> output**

By clicking on “define subcorpus and show texts”, a result similar to the one displayed in graphic 19 will be visible.
In this output, you get a list of all the texts that match the criteria used for defining the subcorpus. In this case, the user gathered all texts that were written in Italian with regard to a specific MERLIN task (not in the screenshot).

The output page specifies the number of texts found ("total hits: 75"). From the ID in the first column, you can understand the target language ("Italian"), the overall rating (differs in the example), and the L1 (in the example: French and Polish).

You can **download** the documents (or only a selection of the subcorpus texts) with or without metadata and with or without target hypotheses in different formats by clicking "download documents".

Also, you can click on **"View learner text and TH"** to get the original text and the target hypothesis 1 or 1 and 2 (an example is shown in graphic 20).
Graphic 20: exemplary output of “view learner text and TH” in <<Define a subcorpus>> output

By clicking on “View learner info and ratings”, metadata for a specific text is displayed (graphic 21 shows the metadata pertaining to the text in graphic 20 above):

Graphic 21: exemplary output of “view learner info and ratings” in <<Define a subcorpus>> output

The subcorpus can be further explored in the Simple or the Advanced search or in the statistics section.
4.2 **Simple search**

In the **Simple search**, it is possible to search for word forms:

The search can be run...
- in the learner text or in the target hypotheses (TH1 or TH2),
- in the entire MERLIN corpus or in a subcorpus that you specified (**Define a subcorpus**)

The simple search is not lemma-based (all word forms of a lexical entry): it only refers to the exact word form you enter. Thus, if you enter, e.g., “abholen” (like in the example below, graphic 23), you will not get results for “abholst” or “hole...ab”.

You can use the virtual keyboard, if needed. Please also be aware of the fact that MERLIN is based on a limited number of tasks which elicit a constrained range of vocabulary when using the simple search.

**Simple search** output

The simple search output gives you the word you looked for in its immediate context (so-called keyword in context, or KWIC):

![Search Interface](image)
By clicking on the key word in context, the full learner text will be displayed. By clicking on <<view learner info and ratings>>, the metadata of the text will be shown. You can also directly download a .pdf file of the task from there and copy the author’s ID if you want to run more specific searches:
This section allows you a more sophisticated access to the MERLIN corpus. It is possible to combine the search for lemmas/words with the search for a variety of annotations that are available in MERLIN.

You can search...

- in the learner text or in the target hypotheses (TH1 or TH2),
- in the entire MERLIN corpus or in a subcorpus that you specified (<<Define a subcorpus>>)

Furthermore, you can decide whether you want to search for 1 or 2 words or lemmas that ...

- are directly adjacent or appear with a specified number of words in between
- belong to specific word class you are interested in (based on automatic part-of-speech (POS) annotation). If you click on the dropdown menu after having selected the target language, you will get a list with POS abbreviations and short explanations (see graphic 26)
Also, you get access to the **manual annotations** available in MERLIN. To that purpose, choose

- **<<feature 1>>** first to specify what category of annotation you are interested in, e.g., grammar, vocabulary, orthography

- As an option, you can further narrow down the search by defining **<<feature detail>>**: Here, you get a list with all **annotation tags** that have been used in a certain annotation category (in the graphic below, you find an example for grammar). If you do not choose a specific tag here, all grammar tags will be displayed if you chose “grammar” in the **<<feature 1>>**.

---

**Graphic 26: Defining POS tags, <<Advanced search>> (detail, German)**

**Graphic 27: Defining <<feature details>> in an <<Advanced search>> (detail, grammar)**
You can execute this procedure for one word/lemma or for a combination of two words/lemmas. Please note that it is not possible to search for learner language features without specifying a word or lemma first.

If you need more information regarding the annotated features, in the <<MERLIN annotations>> section you can consult a list with all tags and examples, learn more about the MERLIN annotation architecture, and download the MERLIN annotation scheme.

<<Advanced search>> output

The <<Advanced search>> section uses the open source search and visualization architecture ANNIS (www.annis-tools.org) which is why its output looks different from what you get in the other MERLIN interface output sections. The following screenshots guide you through the output.

In graphic 28, the lemma “gebären” ("to give birth to someone") is entered in the <<Advanced search>>.

In the output, you find

1. the number of hits and the number of documents with hits (left hand side)
2. the corpus you searched in (bottom left)
3. a button that removes the left hand side of the output which is important to get to a full view of the learner texts
the possibility to get back to the "Advanced search" - your search will not have be saved there

(5) a link to more information ("About this search output"); a new window with information regarding the single tiers of the annotation will open (see chapter 5 of this document)

(6) a link to a list with short explanations of all abbreviations, mainly annotation tags, that are used in the output will open in an extra window ("Abbreviations")

(7) the MERLIN help function

If you hide the left vertical output area which gives you meta information by clicking on , it is easier to look at the learner texts directly:
You see the lemma searched for in its immediate context which you can enlarge to up to 25 tokens on either side:

![Context Enlargement](image)

**Graphic 31: Modify amount of context of a lemma in <<Advanced search>> output**

Please note that you can access automatic annotations, a full view, the transcript, and dependency arcs of the search result. In graphic 32, there is an example in which the automatic annotation is shown in a grid:

![Automatic Annotation](image)
Graphic 32:  View automatic annotations in <<Advanced search>> output

The lines of the output table view are explained in <<about this search output>>.
In this section, some fundamental statistical information is available. You can search in all MERLIN texts or in your choice of texts (“subcorpus”, see section <<Define a subcorpus>>.)

**Graphic 33: Search interface <<Statistics>>**

**Notes of caution for interpretation of statistical measures**

The MERLIN statistical measures are to be interpreted with extreme caution. They can deliver indications and tendencies, but they must not be over-interpreted as evidence for language learning routes. There are different reasons for this, some of which are listed here:

1) MERLIN is a small corpus – simple generalizations are not possible

2) Database for EA1 is different from database for EA2; the latter is extremely small

3) Contrastive analyses should be handled with care. It is not straightforward to compare aggregated measures of different languages because there are some language-dependent tags such as, for example, errors regarding aspect for Czech and Italian, but nor for German.

4) The measures strongly depend on the range of tasks used in MERLIN; this is true especially for the vocabulary and the other EA2 fields

In the <<statistics>> section, there are **four types of measures** available. For any choice you make, you can select multiple annotated features by holding down STRG:
**1. Frequency of annotated features**

Absolute number of annotated features in your database. You can choose one or more feature categories (e.g., grammar, orthography). Then, you can either look for a count of all annotations regarding that category/those categories (e.g., all grammar annotations in German texts) or specify which annotations you are interested in on a more specific level (e.g., all morphological errors in German texts):

```
Graphic 34: Exemplary calculation of absolute frequency (all grammatical errors, German), <<Statistics>>
```

The **output** gives you an overview of the total number of the feature(s) searched for:
If you click on the number of hits (in our example, 5713), you will be directed to the <<Advanced search>> section with direct access to all annotations relevant to your search:

(2) **relative frequency of annotated features**
Relative number of annotated features in your database per sentence or per token. As text length varies considerably throughout the corpus, it is useful to use normalized frequencies, i.e. numbers of occurrences of one or more phenomena with regard to a standardized entity, e.g., like in our case, sentences or tokens. With this function you could, to give an example, calculate the average number of grammatical errors in learner texts that were rated B1 as compared to texts rated B2 (define your subcorpora first).
The procedure for displaying relative frequencies is analogue to the procedure mentioned above for absolute frequencies. Again, from the output on <<Statistics >> you can access all examples in the corpus in the <<Advanced search >> output by clicking on your search result.

(3) error-free learner language
These measures depart from a positive perspective towards learner language in that the percentage of language that does not contain any of the annotated error-based features is displayed. You can calculate the percentage of error-free sentences or tokens with regard to the total number of sentences or tokens. On a more fine-grained level, you can also find out the percentage of language without any grammatical, morphological, or capitalization errors, for example.

(4) complexity measures (German only)
For German, it was possible to include automatically calculated measures of morphological, lexical, and syntactical aspects of complexity in the statistics section. For Italian and Czech, unfortunately, the technical prerequisites were not given.
Complexity is an important aspect of (learner) language that has a close relationship to proficiency. In many studies, it has been shown that complexity is quite clearly distinguishable from accuracy and fluency (so-called ‘CAF’ studies, see bibliography for many references). The measures that MERLIN users have access to stem from research on L2 complexity and readability assessment. They were first implemented by Hancke (2013) and Hancke & Meurers (2013) with regard to MERLIN data (<<documentation>>).
5. Help

There are many documents and functionalities to help you find your way through the MERLIN interface. You can access the help section ( ) from anywhere on the interface.

5.1 User manual

This document, the user manual, is available on the <<help>> page on the MERLIN interface in German, Italian, English, and Czech and will be continuously updated.

5.2 Getting to know MERLIN: video introduction

You can watch two video registrations (each about 30 minutes) in which an overview of the MERLIN project is given (in English, by K. Wisniewski). One introduction is directed towards language teachers, the second one is more appropriate for language testers or textbook authors. The presentations were registered during two workshops that the MERLIN team carried through in Linz in December 2014.
5.3 MERLIN interface navigation: The screencast tutorial

The screencast tutorial is available in English and German and guides you through the main functionalities of the MERLIN interface. The tutorial has two blocks: a basic first part (<<Define a subcorpus>>, <<Simple search>>) enables you to get started by putting together and/or downloading texts / tasks you are interested in and search for words in them. In a second part, explanations on how to run an <<Advanced search>> on lemmas or annotations are given, and the <<statistics>> section is briefly introduced.

5.4 How to apply MERLIN: usage scenarios

In December 2014, the MERLIN team organized multiplier workshops in Linz, Austria, in which the MERLIN interface was introduced with the help of exemplary usage scenarios directed towards language teachers, testers, and trainers. These materials are freely available for download in the <<Using MERLIN>> section, as well, in Czech, Italian, and German. Please be aware of the fact that by the time the workshops were carried through, some MERLIN functionalities had not been implemented yet so that the scenarios might look slightly different from the current status of the MERLIN interface.

5.5 Understanding the <<Advanced search>> output

As mentioned above (chapter 4.3), the <<Advanced search>> uses the open source search and visualization architecture ANNIS. In the search output, in addition to the built-in features of ANNIS, the MERLIN team inserted some help functionalities to make it easier to understand.

Graphic 38: Help in output of exemplary <<Advanced search>>
By clicking on "about this search output", an extra window opens so that you can continue to analyze the search output. All tiers of the Advanced search output grid are explained here (graphic 39).

Graphic 39: "about this search output" help function in "Advanced search" output

Another help option is to open a list with the "Abbreviations" used (graphic 40). The annotation tags with short definitions are displayed in an extra window.
5.6 Glossary

In the general glossary that you find in the help section ( ), terms used on the MERLIN interface are explained, many of them related to annotation.

5.7 Frequently asked questions

Here, questions by users are collected. You find information on what to use MERLIN for, on what to do with your search results, and on the <<Advanced search>> output, for example. This list will be continuously updated.

5.8 Contact us

Do not hesitate to contact the MERLIN team with any question or comment that arises (info@merlin-platform.eu). We are happy to help.
References


Al-Jarf, R. Spelling error corpora in EFL. In the proceedings of the International Conference on Multi Development and Application of Language and Linguistics, National Cheng Kung University, May 5-16 2009, Tainan City, Taiwan, 2009.


Regan, Vera/Howard, Martin/Leme, Isabelle (2009): The Acquisition of Sociolinguistic Competence in a Study Abroad Context.


UIMA Framework: http://uima.apache.org (July 2013)


