The CEFR level descriptions and empirical learner data: MERLIN, a multi-lingual corpus initiative

Katrin Wisniewski, Technical University Dresden
Andrea Abel, European Academy Bozen/Bolzano
Outline

1) The need for illustration of the CEFR levels
2) The Reference Level Descriptions initiative
3) The MERLIN project
4) Applications of MERLIN in Language Teaching
5) Applications of MERLIN in Language Testing
6) Discussion
7) Conclusion/Outlook
1) The need for illustration of the CEFR levels

- in spite of wide distribution of the CEFR, many practitioners do not know what the levels actually mean
- unclear/vague wording, lack of consistence in CEFR, lack of homogeneity
- CEFR text ↔ CEFR scales
- more empirical evidence needed, without analysis of learner language
- applicability across many languages requires general level descriptions ↔ applicability for single languages enhanced by language-specific illustrations

2) The Reference Level Descriptions (RLD) initiative

- Council of Europe initiative: **RLD for National and Regional Languages** (CoE, 2005)
  - level specifications (→ tradition of *Threshold Level*, 1975)
  - purpose: anchor for the development of language programmes
- mostly built on expert intuitions, **top-down approach**, e.g.: *Profile Deutsch* (Glaboniat et al. 2010), A1-C2
- **bottom-up approach** (Green 2013) i.e. RLD based on learner language that has been compiled to a corpus
2) The Reference Level Descriptions (RLD) initiative

Learner corpora

“Computer learner corpora are electronic collections of .... texts produced by foreign or second language learners in a variety of language settings“ (Granger et al. 2002: VII), ”systematic computerized collections of texts produced by language learners” (Nesselhauf, 2005) usually annotated with the help of a standardized system of error tags (Díaz-Negrillo and Domínguez, 2006)

tend to provide meta-information, such as the authors’ L1, age, gender, etc. and other valuable information from all relevant levels of linguistic description

usually a combination of automatic linguistic annotation (e.g. part of speech) & manual annotation (e.g. errors or other features)
2) The Reference Level Descriptions (RLD) initiative

examples of RLD based on learner corpora:

- *Norsk Profil* (Carlsen 2013), ASK learner corpus, B1-B2, restricted access

- *English Profile Project* (Saville/Hawkey 2010), *Cambridge English Profile Corpus*, aims at 10 million words, 20% spoken, 80% written data, A1-C2; English Grammar Profile, English Vocabulary Profile, English Functions Profile; corpus itself is not publicly available

- *Profilo della lingua italiana* (Spinelli/Parizzi 2010), learner corpus (CELI test) A2-B2: 100, corpus itself not publicly available (CD-Rom)
3) The MERLIN project


European Union grant (LLP 518989-LLP-1-2011-1-DE-KA2-KA2MP)

01/2012 – 12/2014

Partners: Technische Universität Dresden (DE) (Lead Partner), European Academy Bolzano (IT), Charles University (CZ), telc GmbH (DE), Berufsförderungsinstitut Oberösterreich (AT), Eberhard-Karls-Universität Tübingen (DE), European Center of Modern Languages - Council of Europe (AT) (associated partner)

**Main aims:**

- develop a freely accessible online platform to **illustrate** the CEFR levels for German, Italian and Czech
- contribute to the **validation** of selected CEFR scales

(cf. Wisniewski et al. 2013)

Katrin Wisniewski & Andrea Abel
3) The MERLIN project: data

- written productions from standardised language tests with strict quality controls
- ~200 texts per (available) CEFR level (N=2470)
- metadata (L1, age, gender ...)
- ~280,000 tokens

<table>
<thead>
<tr>
<th></th>
<th>German</th>
<th>Italian</th>
<th>Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>229</td>
<td>229</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>228</td>
<td>224</td>
<td>131</td>
</tr>
<tr>
<td>B1</td>
<td>231</td>
<td>223</td>
<td>171</td>
</tr>
<tr>
<td>B2</td>
<td>225</td>
<td>222</td>
<td>131</td>
</tr>
<tr>
<td>C1</td>
<td>226</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>1139</td>
<td>898</td>
<td>433</td>
</tr>
</tbody>
</table>
3) The MERLIN project: workflow

1) rater training, re-ratings with CEFR-related analytical rating grid (cf. Table 3, CEFR), 10% double-rated → reliability checks (MFR analyses/CTT)

2) transcription/digitisation (xml mind©) → reliability checks

3) annotation: manual & automatic (ongoing) → reliability checks

4) corpus exploration (queries, visualisation) & corpus statistics (complex indicators)
### 3) The MERLIN project: annotations

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>METHODOLOGY</th>
<th>EXEMPLARY TAGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEFR scales</td>
<td>operationalisation of chapter 5 scales (vocabulary range &amp; control, grammatical accuracy, orthography, coherence &amp; cohesion, sociolinguistic appropriateness)</td>
<td>content jumps, greetings, repetitions, intelligibility</td>
</tr>
</tbody>
</table>
| users (teachers, testers, learners) | • needs analysis (questionnaires & expert interviews)  
• textbook analyses  
• language test analyses | verb mood, verb aspect, incorrect use of prepositions, capitalization errors |
| research           | analyses of research literature - learner corpora, SLA, LT (language-specific & overarching) | REQUESTS  
connectors, text-structural means, reference problems ...  
word order, inflection ...  
semantic/formal lexical errors...  
colloctions, idioms ...  
single variational aspects like double pronouns (ITA) |
| learner texts      | inductive qualitative analyses of 10 texts per level & language             | formality problems, ITA: lexicalised clitics             |
3) The MERLIN project: What the platform will offer

- free online access to all resources, open-source tools
- full texts (with & without annotations), metadata, and tasks
- detailed task descriptions (ALTE Grid for writing)
- CEFR profile for each text (grammatical accuracy, vocabulary range & control, coherence/cohesion, sociolinguistic appropriateness, general linguistic range)
- simple & advanced search options
- sort texts: metadata (e.g. L1, age) – annotations – task level – CEFR level ...
- create & export word lists
- display statistical measures (e.g. of lexical variety, grammatical complexity)
- cross-language learner language analyses
- ...

Katrin Wisniewski & Andrea Abel
3) The MERLIN project: applications

- validation of CEFR scales
- second language acquisition research
- natural language processing research
- language teaching
- language testing
4) Applications of corpora in language teaching

- bottom-up approach to language teaching (McEnery/Xiao 2011)
- usefulness of corpora in language teaching widely acknowledged since Johns (1988), growing meaning ← → teaching itself still largely unaffected (Römer 2006, Chambers et al. 2011)
- learner corpora are especially underrepresented in language teaching (Flowerdew 2012, Granger 2009)
- large native speaker (NS) corpora offer information about real-life frequency: frequency ≠ pedagogical relevance, difficulty (Flowerdew 2012, Meunier 2002) → usefulness of learner corpora
4) Applications of corpora in language teaching

4) MERLIN & syllabus development

- **aim**: help with decisions about selection, presentation, progression of items (Aston 2000, Granger 2002, Römer 2008)

- **MERLIN** ...
  
  - background: almost all language syllabi are related to the CEFR
  
  - MERLIN offers empirical data about typical & relevant milestones/errors in learner language with reference to CEFR levels

  - thus helps to ground expectations about learner achievements on learner language that has been reliably rated according to CEFR levels
4) MERLIN & teaching materials development

- corpora can enhance **authenticity** of teaching materials (Römer 2008)
- **MERLIN** ...
  - provides information on learners’ mastery & variety of use of a multitude of language features on different CEFR levels
  - contributes CEFR-related information to (frequency) (learner) dictionaries, reference works, grammars; e.g., can inform usage notes in learner dictionaries (Granger 2002)
  - provides detailed, CEFR-related information about crucial aspects of language learning, such as learners’ use of formulaic sequences/collocations (Wray/Perkins 2000, Pawley/Syder 1983)
4) Direct applications of corpora in language teaching

- autonomous, learner-centred learning; Johns 1997: „Every student a Sherlock Holmes“
- exploit MERLIN to teach (Leech 1997)
- ‘data-driven learning’ (Johns 1991), discovery learning’ (e.g. Aston 2001)
- ‘three I’s’: Illustration – Interaction – Induction (McEnery/Xiao 2011)
- direct NNS corpus applications are rare, mostly carried through on institutional level (Flowerdew 2012)
4) Direct applications of MERLIN in language teaching

- possible activities include:
  - peer-group error analyses with relation to the CEFR (Mukherjee/ Rohrbach 2006); self-assessment for learners
  - find out if L1, task type, age... influences usage by sorting texts
  - compare interlanguages (Czech/Italian/ German)
  - find out how learners rated higher/lower master specific language phenomena
  - compare MERLIN and native speaker corpus (Granger 2002: CIA, Contrastive Interlanguage Analysis)
5) Applications of MERLIN in Language Testing

- Alderson (1996) first identified potential use \( \leftarrow \rightarrow \) still underexploited (Ball 2001, Barker 2004, Flowerdew 2012), but more common than in Language Teaching

- learner corpora contribute to transparency, consistency, comparability of testing \( \leftarrow \rightarrow \) compiling learner corpora is very expensive; as a rule, corpora are not freely available (e.g. Cambridge ESOL, Barker 2010)
5) Applications of MERLIN in Language Testing

- archiving function
- comparison of test forms (Barker 2006), test revisions
- relate language tests to the CEFR
- CEFR-related item writing, developing assessment materials
- empirically-based rating scale construction (Hawkey/Barker 2004, Fulcher et al. 2011)
- CEFR rating scale validation (Wisniewski 2013, forthcoming)
- future developments include automated essay evaluation; detection of cheating, automatic annotation (NLP) of learner language
6) Discussion

- **generalizability** → MERLIN can only give an *orientation*, not representative
- only a few, mainly qualitative studies about *effectiveness* of corpus use on learning process, generally positive results (Chambers et al. 2011)
- tendency for learner corpora to serve only as *negative evidence* (Leech 1997) in teaching
6) Discussion

- **access** → MERLIN is freely **available**
- **needs orientation**, pedagogic mediation at the level of corpus design and corpus exploitation (Braun 2005) → MERLIN: user needs study regarding annotations & interface design/functionalities
- **corpus technology** as possible obstacle for learners & teachers (Chambers et al., Gilquin et al. 2007) → make benefits clear; graded approach (Chambers et al. 2011) → simple/advanced search options in MERLIN
7) Conclusion/Outlook

- MERLIN bases CEFR use on empirical data, learner-centred approach
- helps to control reification of CEFR scale system (Fulcher/Davidson 2007: 232)
- pilot project design; platform can be extended (→ empirically based „criterial features“ for German, Italian, Czech?)
- “(...) corpora will not only revolutionize the teaching of subjects such as grammar (...), they will also fundamentally change the ways we approach language education, including both what is taught and how it is taught.” (McEnery/Xiao 2012)
- further uses (e.g., second language acquisition research, natural language processing)
Thank you for your attention!

Katrin Wisniewski (katrin.wisniewski@tu-dresden.de)
Andrea Abel (andrea.abel@eurac.edu)
References


ALTE grid for writing tasks, http://www.coe.int/t/dg4/linguistic/manuel1_en.asp


Book editor(s): O’Keeffe, A. and M. McCarthy


Egbert, Paulus and Nakamichi 2002: 111


Katrin Wisniewski & Andrea Abel


Saville/Hawkey 2010


