A Trilingual Learner Corpus illustrating European Reference Levels

Abel, Andrea*; Nicolas, Lionel*; Hana, Jirka♦; Štindlová, Barbora♦; Bykh, Serhiy♦; Meurers, Detmar♦

European Academy Bolzano/Bozen*; Charles University Prague♦; Technical University Liberec♦; University of Tübingen♦

andrea.abel@eurac.edu; lionel.nicolas@eurac.edu;
jirka.hana@gmail.com; barbora.stindlova@tul.cz; detmar.meurers@uni-tuebingen.de; sbykh@sfs.uni-tuebingen.de

LCR 2013 – Bergen, Norway
September 27-29 2013
Outline

1. Introduction
2. Data collection and preparation
3. Creation of the annotation schemata
4. Annotation workflow and technical background
5. Conclusions and outlook
1. Introduction: Aims of the contribution

- to present the MERLIN project
- to detail its current state and explain its latest developments
1. Introduction: The MERLIN project


Background

• Common European Framework of Reference for Languages (CEFR) as THE leading instrument for language teaching and certification and the development of curricula in Europe
• Core of the CEFR: reference levels & descriptors
• Not enough data illustrating the meaning of the CEFR levels in terms of authentic learner data (above all for languages other than English)

(cf. e.g. CoE 2001; englishprofile.org, Spinelli/Parizzi 2010, Carlsen 2013)
1. Introduction: The MERLIN project

Aims of the project:

• to develop a didactically motivated, freely accessible online platform containing authentic learner data illustrating the CEFR levels for German, Italian and Czech

• to create a database containing written learner productions deriving from standardized language tests that have been related in a methodologically rigorous way to the CEFR levels

• to contribute to the validation of the CEFR levels of the languages in question

1. Introduction: The MERLIN project

Target groups:
- teachers, teacher trainers, test developers, text book authors, learner, and educational policy makers

Cooperation project:
- Financing: Lifelong Learning Programme (nr. 518989-LLP-1-2011-1-DE-KA2-KA2MP)
- Duration: 01/2012 – 12/2014
- Project partners: Technische Universität Dresden (DE) (Lead Partner), EURAC (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)
2. Data collection and preparation

Data source:

- standardised tests (sources: telc, DE: German & Italian, UJOP-Charles University Prague, CZ, UNIcert, DE: Czech)
- written learner productions (3-4 tasks per language and level) & meta data (L1, age, gender ...)

<table>
<thead>
<tr>
<th>Level</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>
2. Data collection and preparation

Data preparation – re-ratings:

- Professional raters – rater training
- Re-rating of all tests (ca. 10% double rated) using an own CEFR compliant rating grid:
  - holistic rating scale (global scale: general linguistic range) &
  - analytical rating scales: → rating criteria:
    vocabulary range, vocabulary control, grammatical accuracy, orthographic control, sociolinguistic appropriateness, coherence and cohesion
- Quality control of ratings: test analyses, including Multi-Facet Rasch analyses → high degree of reliability

(cf. e.g. CoE 2001, Alderson 1991)
2. Data collection and preparation

Data preparation – transcription:

• Creation of transcription guidelines (including annotations such as insertions, deletions, unreadable elements, entities, emoticons, images, anonymization, ...)

• On-the-fly inline annotation (see section 4. Annotation workflow and technical background)
2. Data collection and preparation

User modelling:

- Collection of requirements of the MERLIN platform on a content as well as a technical level (questionnaire study and expert interviews)

- Content part:
  - users’ needs concerning the features of learner language that are considered most relevant
  - users’ needs concerning the illustration of CEFR levels

- Technical part:
  - users’ needs for search & display functionalities and concerning the technical environment of the platform
3. Creation of the annotation schemata

a) Linguistic analyses:
   • Identification of indicators to describe aspects of learner language (errors and other linguistic characteristics) as basis for data annotation, data analysis and data query
   • Identification of meaningful indicators for German, Italian and Czech
   • Indicators originating from various sources:
     1. CEFR derived indicators
     2. Experientially derived indicators
     3. Deductively derived indicators
     4. Inductively derived indicators
3. Creation of the annotation schemata

Indicators & sources:

1. CEFR indicators: derived on the basis of the operationalisation of CEFR scale descriptors (chapter 5 scales; selected scales: criteria MERLIN rating grid), e.g. collocations, greetings, content jumps or intelligibility

2. Experientially derived indicators (teachers, testers, text book authors):
   - specific CEFR illustration needs on the basis of the projects questionnaire study and expert interviews, e.g. verbal aspect – Italian/Czech, apostrophe use – Italian/German, incorrect use of prepositions
   - features delivered by text book analyses, e.g. German modal verbs
3. Creation of the annotation schemata

3. Deductively derived indicators: extensive research literature review, such as Second Language Acquisition, Language testing, Corpus linguistics, e.g.

- orthography: punctuation, capitalisation, ...
- grammar: word order, negation, ...
- vocabulary: different aspects of lexical knowledge, particular focus on formulaic sequences, ...
- coherence/cohesion: connectors, use of text structural means, ...
- sociolinguistic appropriateness/pragmatics: addressing, requests, ...

→ tags to be used also for calculation e.g. of coherence/cohesion measures such as connector variety and accuracy, vocabulary measures such as (Advanced) Guiraud’s Index, Lexical Density, percentage of error-free clauses etc.
3. Creation of the annotation schemata

4. Inductively derived indicators: on the basis of the linguistic analyses of performance samples (10 texts per level and language), e.g. register – level of formality, semantic errors, use of formulaic sequences, citations from test task, repetitions

3. Creation of the annotation schemata

Finalisation of the annotation scheme:

- Collection of a noticeable variety of indicators for German, Italian and Czech
- Selection of meaningful indicators for German, Italian and Czech
- Preparation of a harmonised annotation scheme taking both into consideration:
  - common features (e.g. register – level of formality; opening/closing formulas; collocations, idioms; grapheme errors), and
  - language-specific features (e.g. gender/article in German, modal particles in German, reflexive pronouns in Czech, pronoun particles in Italian, lexicalised clitics in Italian)
3. Creation of the annotation schemata

Some details:

- record of: error vs. other linguistic characteristic
- annotation includes a combination of
  - linguistic classification of learner language features (e.g. orthography, grammar) &
  - a „target modification“ (e.g. omission, addition) dimension
- annotation hierarchy – 3 levels: Linguistic field, subfield, special phenomenon (e.g. Coherence/Cohesion, coherence, content jumps)
- tag description (guideline for annotators) & examples
- indication of the indicators source (transparent also for MERLIN end users)
- a detailed manual/documentation: description of the annotation scheme & the procedure

(cf. e.g. Díaz-Negrillo/Fernández-Domínguez 2006, Corder (1993 [1973]))
3. Creation of the annotation schemata

b) Target hypotheses:

- Use of target hypothesis annotation with respect to error annotation: target hypotheses are provided for each learner production to explicitly record the forms on which the annotated interpretations are based (on the basis of the FALKO TH rules – HU Berlin):
  - Target hypothesis I ➔ minimal changes: orthographically and grammatically acceptable sentences
  - Target hypothesis II ➔ changes on semantic and pragmatic level, consideration of the context beyond the single sentence

(cf. e.g. Lüdeling 2008; Falko manual 1.0.1: Recnizeck/Walter et al. 2010)
4. Annotation workflow and technical background

Annotation workflow: the humanist view

Digitising process → Automatic annotation → Manual annotation → Corpus exploration → Corpus statistics

If not good enough ...

Piloting phase

Quality control/reliability checks: e.g. inter annotator agreement

(c.f. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the humanist view

→ MERLIN

- Automatic annotation
- Manual annotation
- Corpus exploration
- Corpus statistics

If not good enough ...

Piloting phase

Quality control/reliability checks: e.g. inter annotator agreement

(c.f. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the humanist view
→ MERLIN

- Automatic annotation
- Manual annotation
- Corpus exploration
- Corpus statistics

If not good enough ...

- e.g. manual transcription, inline annotation, corrections
- e.g. annotation schemata, guidelines, revisions

Piloting phase

Quality control/reliability checks: e.g. inter annotator agreement

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the humanist view

→ MERLIN

e.g. manual transcription, inline annotation, corrections

If not good enough ...

e.g. sentence splitting, tokenisation, lemmatisation, POS-tagging

Manual annotation

Corpus exploration

Corpus statistics

Quality control/reliability checks: e.g. inter annotator agreement

(e.g. Glaznieks/Nicolas/Stemle/Abel/ Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the humanist view

⇒ MERLIN

If not good enough ...

e.g. sentence splitting, tokenisation, lemmatisation, POS-tagging

Manual non inline annotation I:
- orthography and grammar;
- Target hypothesis I (whole text)

Manual non inline annotation II:
- vocabulary, sociolinguistics, pragmatics,
- general features;
- Target hypotheses II (erroneous parts)

Corpus exploration

Corpus statistics

(e.g. manual transcription, inline annotation, corrections)

e.g. annotation schemata, guidelines, revisions

Quality control/reliability checks: e.g. inter annotator agreement

Piloting phase

If not good enough ...

e.g. sentence splitting, tokenisation, lemmatisation, POS-tagging

Manual non inline annotation I:
- orthography and grammar;
- Target hypothesis I (whole text)

Manual non inline annotation II:
- vocabulary, sociolinguistics, pragmatics,
- general features;
- Target hypotheses II (erroneous parts)

Corpus exploration

Corpus statistics

(e.g. Glaznieks/Nicolas/Stemle/Abel/ Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the humanist view

→ MERLIN

- e.g. manual transcription, inline annotation, corrections
- e.g. annotation schemata, guidelines, revisions

Manual non inline annotation I:
- orthography and grammar;
- Target hypothesis I (whole text)

Manual non inline annotation II:
- vocabulary, sociolinguistics, pragmatics,
- general features;
- Target hypotheses II (erroneous parts)

e.g. sentence splitting, tokenisation, lemmatisation, POS-tagging

e.g. corpus queries, visualisation of results

e.g. vocabulary measures, complexity measures

If not good enough …

Quality control/reliability checks: e.g. inter annotator agreement

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the computational linguist view

Digitising process → Conversion → Manual annotation → Corpus exploration → Data storage → Corpus statistics

Automatic annotation

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the computational linguist view

→ MERLIN

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the computational linguist view

→ MERLIN

XML mind for transcriptions

Automatic annotation

Corpus exploration

Data storage

Corpus statistics

Conversion

MMax2, Falko Excel Addin’s for manual annotation

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/ Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the computational linguist view → MERLIN

(UIMA framework - integration of tools for automatic annotation)

- XML mind for transcriptions
- MMax2, Falko Excel Addin’s for manual annotation

Conversion

- Corpus exploration
- Data storage
- Corpus statistics

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the computational linguist view

→ MERLIN

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the computational linguist view
→ MERLIN

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the computational linguist view
→ MERLIN

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
4. Annotation workflow and technical background

Annotation workflow: the computational linguist view

→ MERLIN

- UIMA framework - integration of tools for automatic annotation
- ANNIS for querying & visualisation
- PAULA exchange format for linguistic annotation

XML mind for transcriptions
SaltNPepper conversion framework
MMax2, Falko Excel Addin’s for manual annotation

- e.g. complexity measures for German: Hancke/Vajjala/Meurers 2012

(cf. e.g. Glaznieks/Nicolas/Stemle/Abel/Lyding: in print, Nicolas/Stemle/Glaznieks: submitted)
5. Conclusions and outlook

Very advanced or completed tasks:

- Technical environment and annotation workflow established and implemented
- Data collected, transcribed and checked for all languages
- Inline annotations completed
- Annotation of target hypotheses I layer completed
- Annotation schemata established, piloting concluded

Next steps:

- Annotation of target hypotheses II layer
- Annotation of errors/linguistic features
- Enhancing automatic annotation
- Preparation of the online platform

(cf. MERLIN overview: Wisniewski/Schöne/Nicolas/Vettori/Boyd/Meurers/Abel/Hana: in print)
Conclusions and outlook

Outlook: The MERLIN online platform – content and functionalities:

- contains a richly annotated trilingual learner corpus (German, Italian, Czech)
- offers free online access, open source license for tools & resources
- shows full texts & language test tasks
- is searchable for a wide spectrum of L2-related phenomena (e.g. word order, collocations, greetings)
- can sort texts according CEFR level – tasks – linguistic phenomena
- produces word lists (e.g. showing words of a particular word class according to CEFR levels)
- delivers selected statistical measures (e.g. vocabulary indices, complexity & accuracy measures)
- compares features of learner languages from 3 language families (Germanic, Roman, Slavic)
Thank you for your attention!

Andrea Abel
for the MERLIN-Team

www.merlin-platform.eu

Andrea Abel
Institute for Specialised Communication and Multilingualism
(European Academy Bolzano/Bozen EURAC)
www.eurac.edu/iscm
1. Introduction: The MERLIN project

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proficient User C2</td>
<td>Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.</td>
</tr>
<tr>
<td>Independent User B2</td>
<td>Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.</td>
</tr>
<tr>
<td>Independent User B1</td>
<td>Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.</td>
</tr>
<tr>
<td>Basic User A2</td>
<td>Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.</td>
</tr>
<tr>
<td>Basic User A1</td>
<td>Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.</td>
</tr>
</tbody>
</table>

(CoE 2001)
1. Introduction: The MERLIN project

<table>
<thead>
<tr>
<th>COHERENCE AND COHESION</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
</tr>
<tr>
<td>C1</td>
</tr>
<tr>
<td>B2</td>
</tr>
<tr>
<td>B1</td>
</tr>
<tr>
<td>A2</td>
</tr>
<tr>
<td>A1</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

(CoE 2001)
Extract from the MERLIN rating grid