Using the CEFR in diagnosing writing in a second or foreign language

Ari Huhta
Riikka Ullakonoja
Lea Nieminen.

University of Jyväskylä
J. Charles Alderson
Lancaster University

EALTA 2014 Warwick
Outline

Based on the DIALUKI project 2010-13

(1) How the CEFR was used in the project
(2) Some key analyses and findings
(3) How these findings might inform us about the CEFR levels
DIALUKI - Diagnosing Reading and Writing in a Second or Foreign Language
Research project 2010-2013: www.jyu.fi/dialuki
Interdisciplinary (language testers, SLA & L1 reading researchers, psychologists)
University of Jyväskylä & Lancaster University (Charles Alderson)
International Advisory Board
Research questions
(those that are addressed today)

– To what extent can different L1 and SFL linguistic, cognitive / psycholinguistic and motivation measures predict SFL R/W performance?

– Can performance on those measures be used to characterise different CEFR proficiency levels?

→ basic research on diagnostic assessment, language proficiency, etc
Informants in DIALUKI (Study 1)

Finnish-speaking learners of English as FL (FIN-ENG group)

- Primary school 4th grade
- Lower secondary school, 8th grade
- Gymnasium 2nd year students (“11th grade”)
- ca. 200 students in each group
- relatively homogenous groups (educational background, formal language studies, L1 / FL relationship)

Russian-speaking learners of Finnish as L2 (RUS-FIN group)

- primary school (3-6th grade; N= 186)
- lower secondary school (7-9th grade; N= 78)
- Relatively heterogeneous groups (time of arrival, L1 / L2 relationship)
(1) How did the CEFR influence the design & conduct of the study?

A. One of the RQs directly relates to CEFR

B. It provides frame of reference against which to interpret some of the more detailed, diagnostic findings

C. Selection of reading tests to be used to measure reading comprehension
   - DIALANG: English and Finnish DIALANG reading tests
   - CEFR-linked tasks from the Pearson Test of English General & Pearson’s lower level English reading tests
(1) How did the CEFR influence …

D. Selected CEFR scales were used as rating scales to assess the level of the learners’ writing performances (L1, L2, FL)

Challenge: (most) CEFR scales are not ideal for rating purposes
– few, if any references to errors, problems, limitations
→ few previous attempts to use them unmodified for rating purposes
We used a compilation of 7 specific writing and discourse competence scales from the CEFR:

- Overall written production
- Written interaction
- Correspondence
- Notes, messages, forms
- Creative writing
- Thematic development
- Cohesion and cohesion

Raters gave one overall rating to each script
Each script was double or triple rated
How successful were the ratings? Did the scale work?

- Ratings were analysed with Facets

  - Ratings were mostly successful:
    - no misfitting raters, a limited number of misfitting ratings
    - the unmodified CEFR scale worked
      - there were no disordered levels and all levels were well separated from each other
      - i.e. the raters could distinguish the levels from each other in their ratings

  - on the whole, the placement of the learners on the CEFR levels can be trusted

(2) Some key analyses & findings

A. Modeling correlates / predictors of SFL writing:
   • **Regression models** that show how the **cognitive, linguistic and motivational** tasks predict writing in English as a foreign language and Finnish as a second language

B. Characterising CEFR (English as FL) writing levels
   • in terms of learners’ performance on different linguistic, cognitive and motivational measures
Findings Part A: Modeling predictors of SFL writing

- Latent regression analyses in the structural equation modeling (SEM) framework, using Mplus v7.2

Questions of interest:
- What kind of regression models could be identified (content)? Amount of variance explained?
- To what extent were the models similar across the two language groups? (or across the age groups)
Example:
Psycholinguistic / cognitive tasks as predictors of writing in English as FL and Finnish as L2 (from research on L1 literacy problems)

1) tasks requiring rapid, fluent reading at word level, rapid access to words in memory
2) tasks requiring processing and manipulation of phonemes in non-word / pseudoword tasks (phonological awareness)
3) tasks tapping working memory

- Tasks were presented both in L1 (i.e. Finnish or Russian) and in SFL (i.e. English or Finnish)

ENG-FIN 4th graders vs. RUS-FIN primary group
Rapidly presented words
- say aloud what you saw on the screen
  (this is what the learners saw)
Similarities & differences across the two language groups (FL English / L2 Finnish)

- Different latent factor models were identified for these two groups that are roughly at the same age.

- Finnish / English **foreign language** group: 3 quite strongly correlated factors (= confirmation of our assumptions)
  - (1) fluent word reading in both L1 and FL, (2) phonological efficiency in both L1 and FL, (3) working memory in both L1 and FL

- Russian / Finnish **second language group**: 2 slightly correlated factors
  - (1) first language cognitive skills, (2) second language cognitive skills
  - 3-factor model did not fit at all

- Explained variance was higher for writing in Finnish as L2 than for English as FL (75% vs 42%)
FIN-ENG 4th graders: cognitive tasks as predictors of writing in English as FL

R² = 42.3%
RUS-FIN primary group: cognitive tasks as predictors of writing in Finnish as L2

$R^2 = 75.3\%$
Findings Part B: Characterising CEFR FL writing levels

- Learners were divided into groups (= CEFR levels) based on their FL writing ability
  - rounding fair averages from the Facets analyses

- Group means were compared in the different cognitive, linguistic and motivation variables
  - Analyses of variance (Anova)
  - Post hoc comparisons of the groups
  - (Bootstrapping of confidence intervals)
Questions of interest:

- Which measures distinguish a wide range of proficiency levels?

- Which measures distinguish only specific levels? (e.g. only between A1 and A2 learners)

- Does this interact with age?
  - Which measures distinguish levels across all age groups?
  - Which measures distinguish levels only among the younger learners, which only among the older learners?
Comparison of groups – Step 1
- results of Anova analyses were compiled for each of the age groups

<table>
<thead>
<tr>
<th>FIN-ENG 4th graders</th>
<th>CEFR level in writing in FL English</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLUENT WORD LEVEL READING</td>
<td>&lt;A1</td>
</tr>
<tr>
<td>L1 rapidly presented words (% correct)</td>
<td>47.1</td>
</tr>
<tr>
<td>FL rapidly presented words (% correct)</td>
<td>39.7</td>
</tr>
<tr>
<td>L1 word list reading (words per second)</td>
<td>0.90</td>
</tr>
<tr>
<td>FL word list reading (words per second)</td>
<td>0.95</td>
</tr>
<tr>
<td>L1 rapid naming (seconds per item)</td>
<td>0.94</td>
</tr>
<tr>
<td>FL rapid naming (seconds per item)</td>
<td>1.52</td>
</tr>
</tbody>
</table>
Comparison of groups – Step 2
- combined the results for all 3 FIN-ENG groups

<table>
<thead>
<tr>
<th></th>
<th>&lt; A1</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLUENT WORD LEVEL</td>
<td>≠</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>READING</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 rapidly presented</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>words (% correct)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FL rapidly presented</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>words (% correct)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 word list reading</td>
<td>(4)</td>
<td>(8)</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>(words per second)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FL word list reading</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>(words per second)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 rapid naming</td>
<td>(8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(seconds per item)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FL rapid naming</td>
<td>4</td>
<td>8</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>(seconds per item)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4 = 4th grade
8 = 8th grade
G = gymnasium
Examples of measures that consistently distinguished between several levels:

**FL vocabulary**

Vocabulary Levels Tests (Schmitt, Schmitt & Clapham 2001)

<table>
<thead>
<tr>
<th>FL VOCABULARY</th>
<th>&lt; A1</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 level</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000 level</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000 level</td>
<td>8</td>
<td></td>
<td>8 G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>3000 level</td>
<td>8</td>
<td></td>
<td>8 G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>5000 level</td>
<td>8</td>
<td></td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>AWL</td>
<td></td>
<td></td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>
(3) How might these findings shed light on the CEFR?

A. Modeling of predictors (SEM analyses)
   - Only indirectly and at a rather general level
   - Comparison of cognitive predictors of SFL writing in the two language groups suggests that FL and L2 learners might differ in terms of what characterises their performance
   - Is a foreign language learner’s B1 the same as a second language learner’s B1? .... even if the language in question is the same
(3) How might the findings shed light on the CEFR?

B. Characterising CEFR levels with reference to performance in a range of measures (cognitive, linguistic, motivational)

- Linguistic characteristics of CEFR levels: e.g., vocabulary knowledge in L1 and SFL, linguistic knowledge (segmentation task), L1 reading and writing
  - already some work on this (cf. the SLATE group)

- Cognitive characteristics of CEFR levels: e.g., working memory, word-level reading skills, phonological processing

- Motivational characteristics of CEFR levels
  - new stuff, how useful?
Examples of linguistic characteristics of CEFR levels based on DIALUKI study

Vocabulary profiles for different CEFR FL writing levels based on the Vocabulary Levels Tests (e.g. for our gymnasium students)

<table>
<thead>
<tr>
<th>CEFR level in Writing</th>
<th>n</th>
<th>2000 level (% correct)</th>
<th>3000 level (% correct)</th>
<th>5000 level (% correct)</th>
<th>Academic WordList (% correct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>30</td>
<td>80</td>
<td>62</td>
<td>40</td>
<td>51</td>
</tr>
<tr>
<td>B1</td>
<td>117</td>
<td>93</td>
<td>79</td>
<td>58</td>
<td>69</td>
</tr>
<tr>
<td>B2</td>
<td>36</td>
<td>98</td>
<td>94</td>
<td>78</td>
<td>83</td>
</tr>
<tr>
<td>C1</td>
<td>2</td>
<td>100</td>
<td>98</td>
<td>98</td>
<td>97</td>
</tr>
</tbody>
</table>

CEFR level = rounded fair average from Facets analysis across all raters (2 per script) and writing tasks (3 per learner)
Do learners at different ages have the same FL vocabulary profile when their FL writing level is the same? (8th graders & gymnasium students; 14 vs 17-year-olds)

<table>
<thead>
<tr>
<th>WRITING LEVEL</th>
<th>n</th>
<th>2000 level</th>
<th>3000 level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8th</td>
<td>Gymn.</td>
<td>8th</td>
</tr>
<tr>
<td>A1</td>
<td>36</td>
<td>45</td>
<td>-</td>
</tr>
<tr>
<td>A2</td>
<td>85</td>
<td>73 ↔ 80</td>
<td>45 ↔ 62</td>
</tr>
<tr>
<td>B1</td>
<td>66</td>
<td>83 ↔ 93</td>
<td>63 ↔ 79</td>
</tr>
<tr>
<td>B2</td>
<td>7</td>
<td>84 ↔ 98</td>
<td>72 ↔ 94</td>
</tr>
<tr>
<td>C1</td>
<td>2</td>
<td>100</td>
<td>98</td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Smaller difference 7 – 14 % points
Greater difference 16 – 22 % points
Why did the 8th graders’ and gymnasium students’ vocabulary profiles differ?

- Differences between the raters? (unlikely)
- Two of the three tasks the learners wrote were different
  - 8th: one A2 level, one B1 level Pearson task
  - Gymn: one B1 level, one B2 level Pearson task
  → task level / task difficulty effect?
- Tasks supposed to be at the same level may still differ along several dimensions (→ task / context effect?)
- A 14-year-old’s B1 is not exactly the same as a 17-year-old’s B1? (→ age effect?)
- Interaction between age / task level / nature of the task?
Examples of other cognitive / linguistic / motivation variables that separated learners at different CEFR levels

- FL phoneme deletion: as above
- L1 & FL segmentation:
  - time: B1/B2
  - accuracy: all levels
- L1 spelling error correction: time & accuracy: B1/B2
- L1 reading & writing: all levels
- Instrumental motivation: A2/B1
- Intrinsinc motivation: A1/A2, A2/B1
We can use such findings to characterise the CEFR levels e.g. in terms of vocabulary profiles

For Finnish learners of English (gymnasium students), we can **describe** their CEFR levels in English (as determined by their reading/writing skills) in terms of their English vocabulary knowledge. For example, at B1 level they know:
- over 90% of the 2000 most common word families
- about 80% of the 3000 most common word families
  (and so on for the 5000 level and AWL vocabulary tests)

We can also **predict** learners’ most probable CEFR level if we know their vocabulary profile
- e.g. Finnish learners (in the gymnasium) who know about 80% of the 3000 level words in English are likely to be able to read and write in English at level B1
Other types of characterisations of the CEFR levels may also be possible

We may be able to characterise all or particular CEFR levels with reference to, for example,

– performance on certain cognitive measures (cognitive operations):
  
  • e.g. ”B2 level learners’ working memory capacity allows them to hold almost 5 digits in memory and manipulate them in such a way that they can repeat them accurately backwards, whereas typical B1 level learners can manage close to 4½ digits”
  
  • e.g. ”B2 learners can recognise and manipulate FL phonemes by e.g. leaving out specific phonemes in oral tasks with almost perfect accuracy whereas B1 learners typically achieve 80% accuracy in the same tasks”

– Characterisations need to be made more general, more abstract
Final thoughts

- The CEFR clearly was useful, even necessary for a study like DIALUKI.
- The implications of this kind of basic research for the CEFR are mixed & not entirely clear yet.
  - Vocabulary profiles and profiles of other linguistic elements → probably quite useful.
  - Cognitive and motivational profiles → increase our understanding of language learning & CEFR but practical uses of that information need more work.

- Profiles might be different for:
  - FL and L2 learners
  - Learners at different ages

- More work needed …
Thank you for your attention!

Questions?
Comments?
Backward digit span working memory test in L1 and FL

- all these tasks are taken individually with the help of a laptop computer, headphones and a microphone
- learners did not see this list, they only heard the numbers / digits and they had to respond by repeating backwards the numbers they heard

<table>
<thead>
<tr>
<th></th>
<th>2 - 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5 7 4</td>
</tr>
<tr>
<td>3</td>
<td>7 - 2 - 9 - 6</td>
</tr>
<tr>
<td>4</td>
<td>4 - 1 - 3 - 5 - 7</td>
</tr>
<tr>
<td>5</td>
<td>1 - 6 - 5 - 2 - 9 - 8</td>
</tr>
<tr>
<td>6</td>
<td>8 - 5 - 9 - 2 - 3 - 4 - 2</td>
</tr>
<tr>
<td>7</td>
<td>6 - 9 - 1 - 6 - 3 - 2 - 5 - 8</td>
</tr>
</tbody>
</table>
Rapid reading (aloud) of a list of real FL words – read as many words as you can in one minute

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>be</td>
<td>as</td>
<td>on</td>
<td>at</td>
<td>do</td>
<td>or</td>
<td>if</td>
<td>we</td>
<td>all</td>
<td>for</td>
<td>you</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>house</td>
<td>under</td>
<td>thing</td>
<td>write</td>
<td>tell</td>
<td>hand</td>
<td>both</td>
<td>move</td>
<td>much</td>
<td>back</td>
<td>little</td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>72</td>
<td>73</td>
<td>74</td>
<td>75</td>
<td>76</td>
<td>77</td>
<td>78</td>
<td>79</td>
<td>80</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>because</td>
<td>develop</td>
<td>between</td>
<td>general</td>
<td>another</td>
<td>present</td>
<td>without</td>
<td>problem</td>
<td>however</td>
<td>country</td>
<td>several</td>
</tr>
</tbody>
</table>
**Phoneme deletion task**

- learners did not see these, they only heard the first ‘word’ in each pair
- repetition + deletion
- Example: ”say ’tauk’ … now say ’tauk’ but do not say ’t’ ”

<table>
<thead>
<tr>
<th>L1 or L2 (Finnish)</th>
<th>FL (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Tauk – auk</td>
<td>6. stanseRt – stanset</td>
</tr>
<tr>
<td>2. Hok – oke</td>
<td>7. dockOAn – dockn</td>
</tr>
<tr>
<td>4. gooK – goo</td>
<td>9. driggLE – drigg</td>
</tr>
<tr>
<td>5. ham – hama</td>
<td>10. norCH – nor</td>
</tr>
<tr>
<td>6. pok – poki</td>
<td></td>
</tr>
</tbody>
</table>
Original Vocabulary Levels Test
- a 2000 level example

1 birth
2 dust
3 operation
4 row
5 sport
6 victory

- each level test consists of 10 such sets of words sampled from the particular frequency band (definitions easier than the target words)
Our modified version

1 birth
2 dust
3 operation
4 row
5 sport
6 victory

urheilu ___
voitto ___
syntyminen ___