Linking a computer learner corpus to the CEFR

Cecilie Carlsen, University of Bergen
in collaboration with
Felianka Kaftandjieva, University of Sofia
Plan

1. Why don’t (more) language test developers use computer learner corpora?

2. Linking a learner corpus of Norwegian (ASK) to the CEFR (why and how)

3. A validation study of the illustrative CEFR-scale of coherence and cohesion
"Three heads are better than one"

Language assessment
Second language acquisition
Corpus linguistics
Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance (Granger 2002).
1. Why don’t (more) language test developers use computer learner corpora?
The role of CLC in language assessment
(Alderson, 1996; Taylor and Barker, 2007)

- to provide an empirically-grounded alternative to the intuitions of test constructors in deciding what to test

- to identify typical errors at different proficiency levels

- to identify typical language features and frequencies at different proficiency levels

- to inform decisions about assessment criteria and the development of rating scales
For learner corpora to be useful to language testers

- The proficiency level of texts need to be known
  requirement (often) not met

- The proficiency level of texts need to be reliably assigned
  requirement (often) not met
2. Linking a learner corpus to the CEFR
Corpus design
- 1700 texts
- Automatic grammatical tagger
- Manual error coding
- Corrected versions of learner texts
- Additional NS texts on the same tasks

Text characteristics
- Written texts (timed)
- Authentic exam situation
- Texts from two official tests of Norwegian L2
- Originally two levels of proficiency: intermediate and advanced
- No available aid (i.e. dictionary)
- Different prompts, mainly descriptive/expository at the intermediate level, and expository/argumentative at the advanced level

http://gandalf.aksis.uib.no/ask/

Learner characteristics
- Second language learners
- Learners with 10 different first languages: (German, Dutch, English, Spanish, Russian, Polish, Bosnian-Croatian-Serbian, Albanian, Vietnamese and Somali)
- Searchable background variables: (age, sex, educational background, country of origin, time in Norway, extent of formal instruction of Norwegian, degree of social contact with Norwegians, level of English etc.)
Proficiency levels in ASK

Språkprøven
(Intermediate level test ~B1)

Test i norsk – høyere nivå
(Advanced level test ~B2+)

Proficiency levels in ASK
Linking ASK to the CEFR

...means re-assessing each corpus texts on the CEFR-scale
WHY link ASK to the CEFR?

- Eliminate a potential source of error in ASK-based SLA-research
- Make ASK more interesting for language test developers/researchers
- Allows investigations of
  - what learners actually can (and can’t) do at different CEFR- levels
  - what characterises A2/B1/B2/etc.-level spelling, grammar, vocabulary, discourse etc.
- Link between SLA-studies of inter-language development and levels of proficiency
- Allows collaboration between corpus linguists, SLA-researchers, language testers (SLATE-network)
• Allows an empirical validation of the general level descriptors of the CEFR

• Allows development of *Reference level descriptions for national and regional languages*” (RLD)
How we linked ASK to the CEFR

F. Kaftandjieva’s design

ASK
10 L1s
1700 texts

7 L1s
1222 texts

200 texts
10 raters

511 texts
5 raters

511 texts
5 raters

FEEDBACK

Statistical analysis of rater severity & reliability
F. Kaftandjieva

Level placement, based on mean scores
C. Carlsen

C1/C2
C1
B2/C1
B2
B1/B2
B1
A2/B1
A2
<table>
<thead>
<tr>
<th>Indices</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Range</td>
<td>0</td>
<td>0.91</td>
<td>2</td>
</tr>
<tr>
<td>2. Abs. agreement (%)</td>
<td>40%</td>
<td>64%</td>
<td>100%</td>
</tr>
<tr>
<td>3. Homogeneity index (H)</td>
<td>+0.64</td>
<td>+0.84</td>
<td>+1.00</td>
</tr>
<tr>
<td>4. Inter-rater correlation ($r_{ij}$)</td>
<td>+0.70</td>
<td>+0.82</td>
<td>+0.91</td>
</tr>
<tr>
<td>5. Correlation with the rest ($R_{i-tot}$)</td>
<td>+0.82</td>
<td>+0.90</td>
<td>+0.97</td>
</tr>
<tr>
<td>6. Kappa</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Kendall's W</td>
<td>0.69_{Level1} &amp; 0.68_{Level2}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. ICC (single measures)</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Cronbach's $\alpha$</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. A validation study of a CEFR-scale

Coherence and the use of discourse connectives across CEFR-levels

(Study will be presented in the forthcoming SLATE-publication 2010)
Discourse connectives

• Fraser (1996) defines discourse connectives as elements “which signal a relation between the discourse segment which hosts them, and the prior discourse segment” (p. 190)

• “Linking words” such as: and, for example, in addition, but, despite, however, because, since etc.

• In my study: 36 connectives
<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C2</strong></td>
<td>Can create coherent and cohesive text making <em>full and appropriate use</em> of a variety of organisational patterns and a <em>wide range of cohesive devices</em>.</td>
</tr>
<tr>
<td><strong>C1</strong></td>
<td>Can produce clear, smoothly flowing, well-structured speech, showing <em>controlled use</em> of organisational patterns, <em>connectors and cohesive devices</em>.</td>
</tr>
<tr>
<td><strong>B2</strong></td>
<td>Can use a <em>variety of linking words efficiently</em> to mark clearly the relationships between ideas. Can use a <em>limited number of cohesive devices</em> to link his/her utterances into clear, coherent discourse, though there may be some “jumpiness” in a long contribution.</td>
</tr>
<tr>
<td><strong>B1</strong></td>
<td><em>Can link</em> a series of shorter, discrete simple elements into a connected, linear sequence of points.</td>
</tr>
<tr>
<td><strong>A2</strong></td>
<td>Can use <em>the most frequently occurring connectors</em> to link simple sentences in order to tell a story or describe something as a simple list of points. Can link groups of words with <em>simple connectors</em> like “and”, “but” and “because”.</td>
</tr>
<tr>
<td><strong>A1</strong></td>
<td><em>Can link</em> words or groups of words with <em>the very basic linear connectors</em> like “and” or “then”.</td>
</tr>
</tbody>
</table>

*The Illustrative scale of Coherence and Cohesion, CEFR p. 125., my emphasis*
CEFR-predictions about connectives

• H1: Texts at higher levels contain a *broader range* of different cohesive devices than texts at lower levels

• H2: Texts at higher levels contain more *low-frequency* connectives than texts at lower levels

• H3: Texts at higher levels show a *greater degree of control* of the cohesive devices used than texts at lower levels
Results

• H1: Texts at higher levels contain a broader range of different cohesive devices than texts at lower levels Supported by the data

• H2: Texts at higher levels contain more low-frequency connectives than texts at lower levels Supported by the data

• (H3: Texts at higher levels show a greater degree of control of the cohesive devices used than texts at lower levels) (Supported by the data)
An interesting finding…

• The CEFR predicts that it is only at the B2+ level that a “variety of linking words” are used.

• This study shows that learners even at a B1/B2 level use a range of different connectives, and to an extent which separates them sharply from the lower levels.

• Revision of the scale warranted?
Collaboration is the future

• CLC-design needs insights from the field of language assessment to obtain reliable level assignment of corpus texts

• CLCs with texts reliably placed on proficiency levels may be of great value to language assessment

• Linked to the CEFR, CLCs can also be used in a validation of the CEFR and in the development of RLD

• Allow national and international collaboration between corpus linguists, SLA-researchers and language testers
THANK YOU!