Corpus-aided Error Analysis (CEA) of Accuracy and Proficiency in Learner Finnish

Sisko Bruun, Jarno H. Juntunen, Valtteri Skarpins
University of Oulu, University of Jyväskylä

DATA

The International Corpus of Learner Finnish (ICLF), error-tagged subcorpus (184,089 tokens)
- Source languages: Estonian, Czech, Dutch, Swedish, Chinese
- Proficiency classification based on CEFR (A1–C2)
- Error classification: nine main categories + subclasses

Data in the present study
- A2: 27,196 tokens
- B1: 88,160 tokens
- B2: 54,896 tokens

METHOD

Potential Occasion Analysis (Theuwissen 2015). The POA involves counting errors in relation to the number of occasions that a language producer could potentially make. This method requires both POS and morphologically annotated and error-tagged learner corpora. The former provides information on the overall number of cases studied (potential occurrences of errors) and the error-annotated data the quantitative and qualitative information on the errors made.

A) POTENTIAL OCCASION ANALYSIS:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A2 &gt; B1 &gt; B2</td>
<td>Strong developmental pattern: difference between at least one pair of adjacent proficiency levels</td>
</tr>
<tr>
<td>2. A2 &lt; B1 &gt; B2</td>
<td>Weak developmental pattern: difference between at least one pair of non-adjacent levels</td>
</tr>
<tr>
<td>3. A2 &gt; B1 &lt; B2</td>
<td>Non-progressive trend: shows no area of significant behavioral change</td>
</tr>
</tbody>
</table>

CONCLUSIONS

A) The most frequent pattern is 1 (A2 > B1 > B2) with number of errors steadily and significantly decreasing as proficiency increases.

B) The number of the error categories in which the share of errors decreases is bigger in between B1 and B2:

1. ORTHOGRAPHIC
   - 1A spelling
   - 1B punctuation
   - 1C compounding
   - 3A consonant gradation
   - 3B vowel harmony
   - 5A nominal inflection

2. PHONOLOGICAL
   - 2A quantity
   - 2B diphthong
   - 3A consonant gradation
   - 3B vowel harmony
   - 4A nominal inflection
   - 5A object case and number
   - 5B preposition and number

3. MORPHOSYNTACTIC
   - 5A possessive suffix
   - 5B verb (congruence)
   - 5C subject case and number
   - 5D object case and number
   - 5E prepositional case and number
   - 5F adverbial case and number
   - 5G adverbial and relative
   - 5H preposition (congruence)

5. SYNTACTIC
   - 6A word order
   - 6B non-finite forms and clauses
   - 6C pronouns
   - 6D sentence type

C) In the following error categories the share of errors between A2 and B1 increases:

1. ORTHOGRAPHIC
   - 1A spelling
   - 1B punctuation
   - 5A possessive suffix
   - 5B verb (congruence)
   - 5C subject case and number

D) Non-progressive trend mostly occurs in syntactic errors.

Although the general developmental trend of accuracy in levels A2, B1 and B2 is improvement (Conclusion A, cf. Thewissen 2015: 210), it is noticeable
   - that the specific area where accuracy develops mostly is between levels B1 and B2 (Conclusion B, cf. Thewissen 2015: 184, 272) and
   - that actually between levels A2 and B1 there is sign of regression (Conclusion C, cf. Thewissen 2015: 192)

Some error types, especially syntax errors, show signs of stabilization (Conclusion D, cf. Thewissen 2015: 210).