Academic Writing in English: a corpus-based inquiry into the linguistic characteristics of levels B1-C2

Rebecca Present-Thomas
VU University Amsterdam

Promotors:
John H.A.L. de Jong
Bert Weltens
Background

- The CEFR had been widely adopted as a common tool for describing different levels of proficiency across languages, learners, institutions, etc.

- The functionally motivated illustrative descriptors make the framework and its levels relevant across languages, as different languages utilize different linguistic means of achieving functional goals.

- However, there is concern that existing descriptors are not specific enough to be readily applied in specific settings and for specific target languages. (Alderson, 2007; Fulcher, 2004; Hulstijn, 2007)
Background

• One of the areas in which the existing descriptors are particularly lacking is that of academic writing.

• Given the importance of English as an international language of communication in academia, it is useful to identify relevant linguistic features in English that contribute to or enable language users to achieve the functions described for each level.
Getting from B1 to C2

• B1: Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a breakup in reading.

• C2: Can write clear, highly accurate and smoothly flowing complex texts in an appropriate and effective personal style conveying finer shades of meaning. Can use a logical structure which helps the reader to find significant points.

From the Written Assessment Criteria Grid (Council of Europe, 2009: 187)
Relevant Developmental Indices

- **Lexical**
  - Mean word length (Grant & Ginther, 2000; de Haan & van Esch, 2005)
  - Type/token ratio (Grant & Ginther, 2000; Verspoor et al., 2008)
  - Use of “academic” words (Verspoor et al., 2008; Present-Thomas et al., 2011)
  - Mean word frequency

- **Syntactic**
  - Sentence length/WPS/CPS (Lu, 2010; McNamara et al., 2009; Ortega, 2003)
  - Clause length/WPC (Lu, 2010; Ortega, 2003; Present-Thomas et al., 2013)
  - Left embeddedness/words before main verb (McNamara et al.)
  - Complex NPs/modifiers per NP (Lu, 2010; McNamara et al., 2009, Present-Thomas et al., 2013)

- **Cohesive** (McNamara et al., 2009)
  - Referential cohesion
  - Connectives

Most of these indices can be automatically calculated using Coh-Metrix 3.0.
Current Study

• Which linguistic features are characteristic of academic English texts at levels B1-C2?

• How do the corpora differ from each other when classified by CEF level?
Participants

• Participants: BA students of English Linguistics or English Literature at VU University Amsterdam (Netherlands).

• Students in this population are generally expected to be at level B2 when they enter university and achieve level C2 by the time they leave (Noijons & Kuijper, 2006).
Longitudinal Data Collection

- Data was collected longitudinally over a period of 3 years between September 2010 and May 2013.

- PTEVU: writing tests consisting partially of items from the PTE Academic (~300 words) were administered to incoming students in their first semester and to all students majoring in English at the end of each academic year.

- VUES: longer essays (~1000 words) written in non-timed settings were also collected as (untimed) assignments in the 1st and 2nd year writing courses.
Supplemental Corpora

• Academic writing is a learned construct for both non-native as well as native speakers.

• Therefore, essays were sampled from the British Academic Written English (BAWE) Corpus to see how they compared to the non-native speaker essays.
  – Only essays classified within the English and Linguistics disciplines were included.
Supplemental Corpora

• Academic writing proficiency is a continuous construct, which post-graduates and academics continue to develop well beyond their bachelor studies, internationally published academic text samples from the ICAME corpora (ACE, Brown, FLOB, FROWN, Kolhapur, LOB, WCT) were also included.

• These published texts represent a reasonable point of departure in comparing specific features of learner versus learned writing in order to help highlight developmental trends that might otherwise appear to be insignificant in the learner corpora
  – These texts were assumed to represent (a high) level C2.
CEF Rating

• All essays were rated by at least two trained raters on a continuous scale based on the CEF levels.

• Integers represent boundaries between levels.
• Decimals indicate distance from adjacent levels.

• Ratings were averaged across raters.
CEF Ratings

Mean CEF level per learner corpus

- C2: n = 562, mean = 4.5
- C1: mean = 3.9
- B2: mean = 3.5
- B1: n = 331
- VUES: n = 115
- PTEVU
- BAWE

Grand mean
## CEF Ratings

<table>
<thead>
<tr>
<th></th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTEVU</td>
<td>47</td>
<td>483</td>
<td>32</td>
<td>0</td>
<td>562</td>
</tr>
<tr>
<td>VUES</td>
<td>8</td>
<td>201</td>
<td>116</td>
<td>6</td>
<td>331</td>
</tr>
<tr>
<td>BAWE</td>
<td>0</td>
<td>24</td>
<td>71</td>
<td>20</td>
<td>115</td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>708</td>
<td>219</td>
<td>26</td>
<td>1008</td>
</tr>
</tbody>
</table>
There seems to be a slight increase in word length (lexical specificity) as proficiency increases.
Lexical

There seems to be a slight decrease in type/token ratio (lexical specificity) as proficiency increases.
Lexical

There seems to be a very slight decrease in mean word frequency as proficiency increases.
As proficiency increases, so does the average sentence length.
There is an overall increase in left embeddedness (# words before the main verb) as proficiency increases.
The length of the NP increases, particularly as users progress into the C levels.
More proficient writers seem to use fewer explicit connectives to mark cohesion in their essays.
Cohesion

This represents the degree of overlap of words and ideas across sentences and the entire text. C-level writers tend to use less repetition of words and ideas than B-level writers.
Noun overlap in particular, however, is used less by B1 writers than by B2-C2 writers to mark cohesion across all sentences.
Finally, across adjacent sentences, more proficient writers tend to use noun repetition to mark cohesion even more than they do over the entire text.
Summary of Findings

• On a lexical level, more proficient writers tend to use slightly longer and less frequent words, which is in line with the expectation that they use more lexical specificity or diversity.

• Contrary to expectation, however, they exhibit a lower TTR. However, this may simply be due to the fact that their texts are longer.
Summary of Findings

- Syntactically, more proficient writers use longer sentences consisting of longer noun phrases and a greater number of words before the main verb.
- More proficient academic writers seem to rely on more implicit cohesive markers than do lower proficiency writers, preferring strategic noun overlap to explicit connectives.
Limitations

• Despite rater training and benchmarking, inter-rater reliability was very low (0.35) even when ratings were reduced to a 6-point scale.

• Additionally, the dataset is limited and does not equally represent each of the CEF levels; level C2 learner texts are particularly lacking, as is level B1.
Limitations

• Published (Learned) texts were included in the analysis as a practical attempt to supplement the highest level of academic writing proficiency and highlight trends that reach the highest levels.

• However, the appropriateness of combining (complete) learner essays with published text (samples) can be called into question.
Limitations

- Likewise, as we see in the TTR comparison across CEF levels, there are known task effects that may have confounded the analyses.
- The findings indicate trends across the CEF levels which in most cases do not indicate significant differences.
Selected References

Thank you!
Academic Writing: the construct

• Though it is practical to think about it in terms of levels, language (and writing) proficiency is a continuous construct.

• There can be large within-level differences.
Multiple Linear Regression
(student data only)

```
lm(formula = CEF ~ LDTTRa + WRDFRQa + DESSL + SYNNP + RDL2, data = rated.student.data)
```

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.21963</td>
<td>-0.23902</td>
<td>-0.00738</td>
<td>0.22570</td>
<td>1.73523</td>
</tr>
</tbody>
</table>

Coefficients:

|       | Estimate  | Std. Error | t value | Pr(>|t|) |
|-------|-----------|------------|---------|---------|
| (Intercept) | 5.921152  | 0.490080   | 12.082  | < 2e-16 *** |
| LDTTRa      | -1.532079 | 0.129103   | -11.867 | < 2e-16 *** |
| WRDFRQa     | -0.340805 | 0.169653   | -2.009  | 0.044823 *  |
| DESSL       | 0.022188  | 0.002883   | 7.697   | 3.34e-14 *** |
| SYNNP       | -0.330083 | 0.089210   | -3.700  | 0.000227 *** |
| RDL2        | -0.033297 | 0.003153   | -10.559 | < 2e-16 *** |

---

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3903 on 1002 degrees of freedom
Multiple R-squared: 0.4308, Adjusted R-squared: 0.428
F-statistic: 151.7 on 5 and 1002 DF, p-value: < 2.2e-16

The predictive value of the model is rather low, accounting for 42% of the variance in average text ratings.
Multiple Linear Regression
(Published text samples added)