Automated Assessment of Lexical Diversity and N-grams in Essays at Different Levels of the CEFR

Jeanine Treffers-Daller, Patrick Parslow and Shirley Williams (Reading)
Overview

• Research context
  – Criterial features for different levels of the Common European Framework of Reference
  – Operationalising vocabulary knowledge
• Measures: Lexical Diversity Indices, (POS) N-grams
• Data:
  – 178 essays written for Pearson Test of English Academic
  – Various scores obtained for these essays.
  – data cleaning, lemmatizing
  – Tools: CLAN (MacWhinney, 2000), Gramulator (McCarthy), Speech taggers (Stanford Core NLP tools); Evolutionary Programming, http://www.gene-expression-programming.com/
Overview

• Results
  – Lexical Diversity scores at different levels of the CEFR
  – N-grams
  – Correlations with scores obtained on PTE Academic
  – Evolving models: fitting the best model for the PTE Academic Vocabulary score

• Discussion/Conclusion

• Further research
Research context (1)
The CEFR

- Common European Framework of Reference (2001)
  http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf
- Describes in a comprehensive way what language learners have to learn to do in order to use a language for communication
- What knowledge and skills do they have to develop so as to be able to act effectively.
- CEFR (global descriptors)
Research context (2)

- Few studies relate the functional descriptor scales of the CEFR to the mastery of a number of linguistic features/skills (Kuiken, Vedder and Gilabert 2010)
- Focus on criterial features whose presence or absence can differentiate between these levels (see Bartning, Martin & Vedder 2010 for an overview).
Measures (1): Lexical diversity

- **Lexical diversity**
  
The range or variety of vocabulary, traditionally conceptualised as the number of different words (word types) used in a text or transcript (Malvern, Richards, Chipere and Durán 2004: 192)
types – different words (V)
tokens – total number of words (N)

- **Example 1**
  - We aim to explore the usefulness of the basic list (10 tokens - 9 types)
  - TTR = 9/10 = 0.9

- **Example 2**
  - We aim to explore the usefulness of the basic list in measuring the richness of the vocabulary of the informants (20 tokens, 14 types).
  - TTR: 14/20 = 0.7
Measures of lexical diversity

- TTR Type – Token – ratio (types / tokens)
- Index of Guiraud – (types/SQRT tokens) (Guiraud 1954)
- D (D as a parameter of the ‘TTR curve’) (Malvern and Richards 1997; Malvern, Richards, Chipere and Durán 2004).
- HDD (McCarthy and Jarvis 2007)
- MTLD (McCarthy 2005; McCarthy and Jarvis 2010)
HD-D (Jarvis & McCarthy 2010)

- HD-D calculates, for each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 words drawn from the text (Jarvis and McCarthy 2010: 383).
- Based on the hypergeometric distribution (Wu 1993)
MTLD (McCarthty 2005)

• This measure is calculated as the mean length of sequential word strings in a text that maintain a given TTR value (0.720) (see McCarthy and Jarvis, 2010).

• Example:

• of (1.00) the (1.00) people (1.00) by (1.00) the (.800) people (.667) |||FACTORS = FACTORS = 1||| for (1.00) the (1.00) people (1.00) . . .

• MTLD is obtained by dividing the total number of words by the total number of factors. Thus, if the text is 360 words long and there are 4 factors, the MTLD value is 90.
Measures (2): N-grams

- An **n-gram** is a set of \( n \) adjacent tokens from a body of text. N-grams can be
  - common phrases “every cloud has a silver lining” (6-gram)
  - words which commonly occur next to each other “if we look at the” (5-gram)
  - part of a larger n-gram “we look at” (3-gram)
- The distribution of n-grams sizes in a corpus tends to follow an exponential decay e.g. \( freq = W_c (1 - e^{-n}) \), where \( W_c \) = token count, and \( n = n \)-gram size.
Research questions

1) Which range of scores on different measures of lexical diversity are typically found in essays written by learners at B1, B2, C1 and C2 of the CEFR?

2) Which patterns of n-gram use can discriminate between different levels of the CEFR?

3) To what extent do lexical diversity scores and n-gram usage correlate with scores obtained on the PTE Academic?

4) Which form of lemmatization is most appropriate for the analysis of lexical diversity?
Methods (1)
Pearson Test of English Academic

- Computer-based international English language test
- Designed to measure language competence according to the principles of the CEFR and to address specifically language competencies in the range from upper B1 to lower C2.
- The score report provides three types of scores:
  - Overall Score
  - Scores for Communicative Skills (i.e. Listening, Reading, Speaking and Writing)
  - Scores for Enabling Skills (i.e. Grammar, Oral Fluency, Pronunciation, Spelling, Vocabulary and Written Discourse).

The score scale ranges from 10 to 90.
• Alignment of PTE Academic Overall score and the CEFR
## Methods (2)

### Essays (n = 178)

<table>
<thead>
<tr>
<th>CEFR level</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>50</td>
<td>50</td>
<td>49</td>
<td>29</td>
</tr>
</tbody>
</table>

- **PTE Academic scores**
  - Overall score
  - Writing score
  - Vocabulary score
  - Scores range from 10-90

- Two different prompts (difference n.s.)

- Text length: 187-357 words (M=250, SD = 35.72)

- Corpus size: 44,445 tokens
Methods (3)

- Transcription in CHAT, analysis with CLAN (MacWhinney 2000)
- Data cleaning: removal of typos, proper names, numbers
- What counts as one type?
  - **No lemmatization**: work, works, working, worked (all different types)
  - **Lemmatization 1**: work, works, working, worked (each lemma = one type; disambiguation of homographs)
  - **Lemmatization 2**: work, works, working, worked, workable, worker (all members of word family are one type)
- Morphosyntactic coding on mor tier
  
  prep|to   inf|to
  n|individual-PL  adj|individual
Examples of lemmatization

- **Original**: a concluding paragraph highlighting the main issues that have been investigated

- **Lemmatization 1**
  - det|a v|conclude-PROG n|paragraph v|highlight-PROG
det|the adj|main n|issue-PL rel|that aux|have aux|be&PERF
v|investigate-PERF.

- **Lemmatization 2**
  - a conclude paragraph **highlight** the main issue **that have be** investigate
# Results (1): Lexical diversity

<table>
<thead>
<tr>
<th>Method</th>
<th>F(3, 174)</th>
<th>Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>D (no lemmatization)</td>
<td>6.90</td>
<td>.110</td>
</tr>
<tr>
<td>D (lemmatization 1)</td>
<td>6.73</td>
<td>.104</td>
</tr>
<tr>
<td>D (lemmatization 2)</td>
<td>5.57</td>
<td>.088</td>
</tr>
<tr>
<td>HDD (no lemmatization)</td>
<td>8.78</td>
<td>.132</td>
</tr>
<tr>
<td>HDD (lemmatization 2)</td>
<td>8.09</td>
<td>.122</td>
</tr>
<tr>
<td>MTLD (no lemmatization)</td>
<td>9.88</td>
<td>.145</td>
</tr>
<tr>
<td>MTLD (lemmatization 2)</td>
<td>6.31</td>
<td>.098</td>
</tr>
<tr>
<td>Guiraud (no lemmatization)</td>
<td>20.5</td>
<td>.260</td>
</tr>
<tr>
<td>Guiraud (lemmatization 1)</td>
<td>28.09</td>
<td>.326</td>
</tr>
<tr>
<td>Guiraud (lemmatization 2)</td>
<td>24.29</td>
<td>.295</td>
</tr>
</tbody>
</table>
Results (1): lexical diversity

F(3, 174)=28.09, p<.001

B1 significantly different from all other groups, B2 significantly different from C2.
## Correlations of lexical diversity measures with PTE Academic scores

<table>
<thead>
<tr>
<th>LD</th>
<th>Writing score</th>
<th>Vocabulary score</th>
<th>Overall score</th>
</tr>
</thead>
<tbody>
<tr>
<td>D lemmatized 1</td>
<td>.164*</td>
<td>.184*</td>
<td>.151*</td>
</tr>
<tr>
<td>D lemmatized 2</td>
<td>.175*</td>
<td>.184*</td>
<td>.159*</td>
</tr>
<tr>
<td>HDD (not lemmatized)</td>
<td>.245**</td>
<td>.242**</td>
<td>.231**</td>
</tr>
<tr>
<td>HDD (lemmatized 2)</td>
<td>.224**</td>
<td>.214**</td>
<td>.224**</td>
</tr>
<tr>
<td>MTLD (not lemmatized)</td>
<td>.237**</td>
<td>.248**</td>
<td>.209**</td>
</tr>
<tr>
<td>MTLD (lemmatized 2)</td>
<td>.179**</td>
<td>.191**</td>
<td>.157**</td>
</tr>
<tr>
<td>Guiraud 1</td>
<td>.359**</td>
<td>.400**</td>
<td>.382**</td>
</tr>
<tr>
<td>Guiraud 2</td>
<td>.358**</td>
<td>.396**</td>
<td>.383**</td>
</tr>
</tbody>
</table>
LD and vocab score (text length fixed at 200 words), n=175

<table>
<thead>
<tr>
<th>Metric</th>
<th>F(3, 171)</th>
<th>Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTLD_200 (lemmatization 2)</td>
<td>8.711**</td>
<td>.133</td>
</tr>
<tr>
<td>HDD_200 (lemmatization 2)</td>
<td>6.901**</td>
<td>.108</td>
</tr>
<tr>
<td>D_200 (lemmatization 1)</td>
<td>4.942**</td>
<td>.080</td>
</tr>
<tr>
<td>D_200 (lemmatization 2)</td>
<td>5.625**</td>
<td>.090</td>
</tr>
<tr>
<td>Guiraud_200 (lemmatization 1)</td>
<td>15.084**</td>
<td>.209</td>
</tr>
<tr>
<td>Guiraud_200 (lemmatization 2)</td>
<td>17.081**</td>
<td>.231</td>
</tr>
<tr>
<td>TTR_200 (lemmatization 1)</td>
<td>14.928**</td>
<td>.208</td>
</tr>
<tr>
<td>TTR_200 (lemmatization 2)</td>
<td>17.446**</td>
<td>.234</td>
</tr>
</tbody>
</table>
## Correlations between LD (text length = 200) and Pearson scores, n= 175

<table>
<thead>
<tr>
<th>Metric</th>
<th>Writing score</th>
<th>Overall score</th>
<th>Vocab score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTLD_200</td>
<td>.215**</td>
<td>.203**</td>
<td>.223**</td>
</tr>
<tr>
<td>HDD_200</td>
<td>.219**</td>
<td>.208**</td>
<td>.228**</td>
</tr>
<tr>
<td>D_200 (lemmatized 1)</td>
<td>.156*</td>
<td>.144 (ns)</td>
<td>.179*</td>
</tr>
<tr>
<td>D_200 (lemmatized 2)</td>
<td>.160*</td>
<td>.144 (ns)</td>
<td>.182*</td>
</tr>
<tr>
<td>Guiraud_200 (lemmatized 1)</td>
<td>.321**</td>
<td>.327**</td>
<td>.377**</td>
</tr>
<tr>
<td>Guiraud_200 (lemmatized 2)</td>
<td>.314**</td>
<td>.321**</td>
<td>.382**</td>
</tr>
<tr>
<td>TTR_200 (lemmatized 1)</td>
<td>.323**</td>
<td>.335**</td>
<td>.384**</td>
</tr>
<tr>
<td>TTR_200 (lemmatized 2)</td>
<td>.319**</td>
<td>.331**</td>
<td>.391**</td>
</tr>
</tbody>
</table>
## Results (2) N-grams and CEFR levels, n=178 (ANOVA)

<table>
<thead>
<tr>
<th>N-gram</th>
<th>F(3, 174)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=1</td>
<td>10.887</td>
<td>p = .000</td>
</tr>
<tr>
<td>N=2</td>
<td>2.629</td>
<td>p = .052</td>
</tr>
<tr>
<td>N=3</td>
<td>2.449</td>
<td>p = .065</td>
</tr>
<tr>
<td>N=4</td>
<td>1.323</td>
<td>p = .268</td>
</tr>
<tr>
<td>N=5</td>
<td>.412</td>
<td>p = .745</td>
</tr>
<tr>
<td>N=6</td>
<td>.562</td>
<td>p = .641</td>
</tr>
</tbody>
</table>
N-gram analyses

1 gram frequency by CEFR grade

Frequency of use

b1  b2  c1  c2
N-gram analyses

2 gram frequency by CEFR grade
Indicative Correlations of N-gram counts with PTE Academic scores

<table>
<thead>
<tr>
<th>Metric</th>
<th>Overall score</th>
<th>Writing score</th>
<th>Vocabulary score</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1 count</td>
<td>.169</td>
<td>.163</td>
<td>.142</td>
</tr>
<tr>
<td>N2 count</td>
<td>-.033</td>
<td>.032</td>
<td>.036</td>
</tr>
<tr>
<td>N3 count</td>
<td>-.081</td>
<td>-.057</td>
<td>.009</td>
</tr>
<tr>
<td>N4 count</td>
<td>-.068</td>
<td>-.039</td>
<td>.018</td>
</tr>
<tr>
<td>N5 count</td>
<td>-.018</td>
<td>.001</td>
<td>.043</td>
</tr>
<tr>
<td>N6 count</td>
<td>.027</td>
<td>.035</td>
<td>.058</td>
</tr>
<tr>
<td>Spelling errors*</td>
<td>-.369</td>
<td>-.371</td>
<td>-.353</td>
</tr>
</tbody>
</table>
POS tagging & N-grams

- Part of Speech tagging was performed using the Stanford Core NLP tools
- POS N-grams were determined for the POS tags
  - e.g.
  - “The International Organization for Standardization”
  - <DT NNP NNP IN NNP> (5gram of POS tags)
  - and
  - “We explain the issues of global warming”
  - <PRP VBP DT NNS IN JJ NN> (7gram of POS tags)

(Penn Treebank II Tags)
Evolutionary programming & Data handling

- An evolutionary programming technique was used to find functions which ‘learned’ to match the data.
  - Initial attempts to match the CEFR level (overall score)
  - Later attempts to model the Vocab Score from Pearson
- A range of metrics were used as inputs (and more to be evaluated)
- e.g. using the count of n-grams used (n=1 to 6) and the count of different POSs used in each essay
Observations on data

- Small data set for machine learning purposes
- ‘Noisy’ data
- Hypothesis that n-gram use and/or POSs are related to the Vocab score
- Data subdivided to allow for training and testing sets
  - 3 subsamples for each CEFR grade
  - Can be recombined into $2 \times 3^4 = 162$ training/testing sets
Evolving models

- Correlation 0.46, RMSE 17.9

(produced using GeneXproTools)
Testing evolved models

- Correlation 0.43, RMSE 19.9
  (produced using GeneXproTools)
Extracting knowledge

- Although this evolved model is not yet complete
  - We can extract knowledge about the functions it is learning
  
  - In this case, the ‘program’ which is evolved consists of 16 sub-expressions (represented above as ‘trees’)

- Functions used:
  - + - * / sqrt exp ln \(1/x\) \(x^2\) max, min, and, or (etc)
  
  - Actually learns better if allowed to use sin, cos
Conclusion + discussion

- Lexical diversity measures useful for discriminating between levels of CEFR – in conjunction with other measures
- Discrimination between B1 and other levels promising
- Guiraud more powerful than other measures in discriminating between levels of CEFR.
- TTR strongest – if word length is kept constant.
- Lemmatization 2 (type = word family) discriminates better between groups than lemmatization 1 (type = lemma + disambiguation of homographs)
- Moderate correlations with Pearson’s scores, in particular the vocabulary score.
Conclusion + discussion (n-grams)

- N-gram analysis needs refining (work on clean data, lemmatized data)
- POS n-grams potentially more promising than n-grams based on fixed sequences of words.
- Evolutionary program method needs refining - further repetitions to provide better fit with vocab scores
Further work (Lexical Diversity)

- **What is a type?**
  - Type = lemma
    - work, worker, workable (different types)
  - Type = word family
    - Work, worker, (un)workable (one type)

- **Homographs (noun – verb)**
  - I’m *meeting* with John at 2pm (verb)
  - I have a *meeting* with John at 2pm (noun)

- **Homographs in function words**
  - I know *that* ... (subordinate conjunction)
  - The book *that* you bought... (relative pronoun)
  - *That* book is... (determiner)
Further work (Lexical diversity)

• Standards for different task types
  – Narrative
  – Essays

• Standards for different languages
  – Is D score of 50 comparable in French and English?
    • Differences in inflectional/derivational morphology
  – How do LD scores correlate with tests of productive vocabulary (Nation’s levels test)?
Further work (Lexical diversity)

- Standards for different modalities
  - Written (formal/informal)
  - Spoken (formal/informal)

- Standards for data treatment
  - Data cleaning (spelling errors, incorrect usage of words, acronyms, proper names, numbers)

- Standards for data analysis
  - Black box programs?
Further work (evolutionary programming)

- Improve modelling with evolutionary programming techniques
- N-gram processing of lemmatized data
- Further investigation of potential metrics, and combinations of metrics
Acknowledgements

- Pearson for funding of the project
- Kirsten Ackermann and Veronica Benigno for their advice in working with the Pearson data set.

- Thank you very much!
N-gram analyses

2 gram frequency by CEFR grade

Frequency of use

b1  b2  c1  c2
N-gram analyses

3 gram frequency by CEFR grade
N-gram analyses

4-gram frequency by CEFR grade
N-gram analyses

5 gram frequency by CEFR grade
N-gram analyses

6 gram frequency by CEFR grade

b1  b2  c1  c2
<table>
<thead>
<tr>
<th>CEFR</th>
<th></th>
<th>types_0</th>
<th>tokens_0</th>
<th>types_1</th>
<th>Tokens_1</th>
<th>Types_2</th>
<th>Tokens_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>Mean</td>
<td>116.8200</td>
<td>225.6200</td>
<td>116.9400</td>
<td>225.62</td>
<td>108.94</td>
<td>227.42</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>B2</td>
<td>Mean</td>
<td>139.4200</td>
<td>253.8800</td>
<td>138.3600</td>
<td>253.78</td>
<td>128.36</td>
<td>254.98</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>16.41439</td>
<td>32.06667</td>
<td>16.91124</td>
<td>32.014</td>
<td>15.759</td>
<td>32.269</td>
</tr>
<tr>
<td>C1</td>
<td>Mean</td>
<td>141.9184</td>
<td>256.1837</td>
<td>142.0000</td>
<td>256.22</td>
<td>131.53</td>
<td>257.12</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>C2</td>
<td>Mean</td>
<td>152.5172</td>
<td>273.0000</td>
<td>152.6552</td>
<td>272.90</td>
<td>141.07</td>
<td>274.00</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>15.85339</td>
<td>29.34767</td>
<td>15.56846</td>
<td>29.369</td>
<td>14.871</td>
<td>29.245</td>
</tr>
<tr>
<td>Total</td>
<td>Mean</td>
<td>135.8933</td>
<td>249.6910</td>
<td>135.6742</td>
<td>249.66</td>
<td>125.85</td>
<td>250.93</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>178</td>
<td>178</td>
<td>178</td>
<td>178</td>
<td>178</td>
<td>178</td>
</tr>
</tbody>
</table>