Bundles in learner corpora: what a type and token analysis can reveal

Deise Prina Dutra- UFMG  
deisepdutra@gmail.com

Barbara Malveira Orfano-UFSJ  
blmalveira@yahoo.com.br

Tony Berber–Sardinha-PUC-SP  
tony@pucsp.br
Acknowledgment

• Faculdade de Letras da UFMG
  – LEEL
• Centro de Extensão da UFSJ
• PUC-SP
• FAPESP
Introduction

- Corpus Linguistics (CL) has valued the investigation of group of words rather then words in isolation
  - Collocations (Sinclair 1991)
- Studies have concentrated on lexical bundles in a variety of contexts
  - in business contexts – genre based analysis of business report (Berber Sardinha 2003);
  - in the university – oral and written discourse - (Biber et al. 2004; 2006; 2009);
  - in different disciplines– electric engineering, biology, administration, applied linguistics (Hyland 2008);
  - in academia, where Simpson-Vlach and Ellis (2010) propose a list of the most commonly used bundles in academic registers.
Lexical Bundles

• simply sequences of word forms that commonly go together in natural discourse (Biber et al. 1999: 990)
  – *in terms of the*
  – *a list of*
  – *the fact that*
  – *it has been argued that*
  – *to a certain extent*
  – *my point of view*
Research on lexical bundles

• Biber et al. (2004)
  – Frequency approach
  – Classroom teaching and textbooks
  – Structural patterns and function
    – Three major functional categories
      » Referential expressions
      » Stance expressions
      » Discourse organizing functions

• Simpson-Vlach e Ellis (2010)
  – oral and written corpora
  – MICASE + BNC (oral academic part)
  – Hyland corpus (2004) + BNC files (various academic subjects)
  – Academic Formulas List (AFL)- 435 lexical bundles
Aims

- to discuss the relevance of analyzing and contrasting types and tokens of bundles produced by native and non-native speakers in argumentative essays;

- to highlight the differences among the corpora as far as stance expressions are concerned;

- to detect if these differences are mainly structural or related to frequency within a specific function.
Data - Essays

- **LOCNESS (Louvain Corpus of Native English Essays)**
  - 324,006 words
  - written language
  - American and British university students

- **ICLE (International Corpus of Learner English)**
  - 3.7 million words (Granger et al. 2009)
  - written language
  - 16 subcorpora (Japan, China, Italy, Finland ...)

- **Br-ICLE (Berber Sardinha 2001)**
  - In 2009-> 159,000 words (aim 200,000 words)
    - CABrl (Corpus de Aprendizes Brasileiros de Inglês – UFMG)

- **Total** – 4,251,714 words
Methodology

- bundles of 4 words were extracted from each corpus with scripts specially developed for our research project;
- the bundles were categorized manually and automatically according to the AFL framework
  - 3 major categories: referential expressions, stance expressions and discourse organizing functions - 18 specific subcategories
- the most frequent categories in each corpora were identified and isolated and we detected the differences in terms of types of bundles across the broad categories (\(>= 20 \text{ wpm}\));
- token frequency analysis was done to investigate the extent to which they could reveal significant differences among the subcategories;
- we ran statistical tests to identify differences within each category;
- concordance lines for the most frequent bundles in each corpora were generated in order to identify differences in use across the 3 datasets.
A = referential expressions
B = stance expressions
C = discourse organizing expressions

The graph shows the count of each type of expression across different corpora: BRIC, ICLE, and LOCN. The bars indicate the frequency of each type of expression in the respective corpora.
# Chi-square Test

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymp.Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>17.126</td>
<td>4</td>
<td>0.002</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>17.508</td>
<td>4</td>
<td>0.002</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>676</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Chi-square test

all subcategories

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>79.624</td>
<td>34</td>
<td>0.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>23.112</td>
<td>34</td>
<td>0.000</td>
</tr>
<tr>
<td>N Valid Cases</td>
<td>676</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## SUB-CATEGORY X CORPUS
(TOKEN FREQUENCY)

<table>
<thead>
<tr>
<th>Sub-Category</th>
<th>LOCNESS</th>
<th>ICLE</th>
<th>BRICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>raw</td>
<td>wpm</td>
<td>raw</td>
</tr>
<tr>
<td><strong>B1 Hedges</strong></td>
<td>33</td>
<td>101.851</td>
<td>104</td>
</tr>
<tr>
<td><strong>B2 Epistemic stance</strong></td>
<td>83</td>
<td>255.992</td>
<td>2128</td>
</tr>
<tr>
<td><strong>B3 Obligation and directives</strong></td>
<td>75</td>
<td>231.478</td>
<td>1485</td>
</tr>
<tr>
<td><strong>B4 Expressions ability and possibility</strong></td>
<td>97</td>
<td>299.379</td>
<td>1252</td>
</tr>
<tr>
<td><strong>B5 Evaluation</strong></td>
<td>129</td>
<td>370.364</td>
<td>2485</td>
</tr>
<tr>
<td><strong>B6 Intention, volition and prediction</strong></td>
<td>32</td>
<td>98.763</td>
<td>748</td>
</tr>
<tr>
<td>LOCNESS</td>
<td>ICLE</td>
<td>BR-ICLE</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>-----------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>B1</strong> to a certain extent could be used to</td>
<td>is a kind of</td>
<td>is a kind of</td>
<td></td>
</tr>
<tr>
<td>can be seen to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I think that the</td>
<td>I think it is</td>
<td>it has been argued that</td>
<td></td>
</tr>
<tr>
<td>I feel that the</td>
<td>I do not think</td>
<td>some people think that</td>
<td></td>
</tr>
<tr>
<td>can be seen as</td>
<td>I think that the my point of view</td>
<td>think that it is</td>
<td></td>
</tr>
<tr>
<td>is seen to be</td>
<td>seems to be a</td>
<td>my point of view</td>
<td></td>
</tr>
<tr>
<td><strong>B2</strong> is shown to be</td>
<td>I think it is</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I think that the</td>
<td>I do not think</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel that the</td>
<td>I think that the my point of view</td>
<td></td>
<td></td>
</tr>
<tr>
<td>can be seen as</td>
<td>seems to be a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is seen to be</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B3</strong> would have to be</td>
<td>do not want to</td>
<td>what they want to</td>
<td></td>
</tr>
<tr>
<td>it should not be</td>
<td>they do not have to</td>
<td>you do not have</td>
<td></td>
</tr>
<tr>
<td>should be able to</td>
<td>think that it is</td>
<td>we need to be</td>
<td></td>
</tr>
<tr>
<td>should not be allowed</td>
<td>do not have to</td>
<td>do not need to</td>
<td></td>
</tr>
<tr>
<td>should be allowed to</td>
<td>should be able to</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bundle Structure

- Preposition +NP – *to a certain extent*
- Passive- *can be seen to*
- (NP)+ V + *that*-clause – *think that it is*
- VP (Modal + V) – *would have to*
- Copula *be* + NP or AdjP – *is a kind of*
- Antecipatory *it* + VP/AdjP – *it should not be*
Register appropriateness

- **Written vs Spoken**
  - Hedging (cautious language)
    - LOCNESS
      - *to a certain extent / could be use to / can be used to*
    - ICLE and Br-ICLE
      - *is a kind of*
- **Participant-oriented (reader or writer oriented)**
  - Epistemic
    - LOCNESS
      - *is shown to be / can be seen as / is seen to be*
      - *I think that the / I feel that the*
    - ICLE and Br-ICLE
      - *I think it is / some people think that / my point of view*
      - *it has been argued that*
Chi-square test
Stance expressions

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square Valid Cases</td>
<td>8.742</td>
<td>10</td>
<td>0.557</td>
</tr>
</tbody>
</table>
Normalized token frequency

- Hedging
- Epistemic
- Obligation
- Directives
- Ability
- possibility
- Evaluation
- Intention
- Volition
- Prediction
<table>
<thead>
<tr>
<th>B3</th>
<th>would have to be</th>
<th>do not want to</th>
<th>what they want to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>it should not be</td>
<td>they do not have to</td>
<td>you do not have</td>
</tr>
<tr>
<td></td>
<td>should be able to</td>
<td>think that it is</td>
<td>we need to be</td>
</tr>
<tr>
<td></td>
<td>should not be allowed</td>
<td>do not have to</td>
<td>do not need to</td>
</tr>
<tr>
<td></td>
<td>should be allowed to</td>
<td>should be able to</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion
Br-ICLE

• Types
  • Less diverse use of stance bundles

• Tokens
  • More personal
    • bundle structure
      • fewer antecipatory *it* and passive structures
  • Directive and obligation
    • Participant-oriented
      • fewer hedging bundles
      • instead there is overuse of bundles that carry an
        overstating tone

• Lexical bundle studies
  • Token analysis complements type analysis helping to
    describe different corpora even when there are no
    statistically significant differences.
Future actions

• Classify more bundles - >10 wpm
  – Improve automatic bundle classification

• Bundle analyzer
  – Make it available to
    • Teachers
    • Students

• Add to the bundle analysis
  – Readability measures
Bibliography


CHEN, Y.; BAKER, P. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*. June 2010, Volume 14, Number 2 pp. 30–49


DUTRA, D. P.; BERBER-SARDINHA, T. Pacotes lexicais em corpora de aprendizes. (in press)


OLIVEIRA, M.; DUTRA, D. Pacotes lexicais ou palavras isoladas? Organizadores discursivos em corpora de aprendizes e de falantes nativos. 2011


Thank you!

Deise Prina Dutra- UFMG
deisepdutra@gmail.com

Barbara Malveira Orfano-UFSJ
bmalveira@yahoo.com.br

Tony Berber–Sardinha-PUC-SP
tony@pucsp.br