Positional variation of phrase-frames in a new corpus of proficient student writing

Ute Römer & Matthew Brook O‘Donnell

AACL Edmonton, Canada – 9 October 2009

www.elicorpora.info
Presentation outline

1. Introduction to MICUSP (Michigan Corpus of Upper–level Student Papers): Composition and markup

2. Positional variation of phrase–frames in MICUSP
   2.1 Method
   2.2 Select results

3. Conclusion

Ute Römer (uroemer@umich.edu)
Matthew Brook O’Donnell (mbod@umich.edu)
1. Introduction to MICUSP: Composition and markup

**MICUSP background**

- Rather extensive research on academic writing produced by experts and by learners (CL and EAP)
- Gap: advanced but unpublished academic writing by graduate–level university students
- Why? Difficulty of accessing unpublished academic writing—especially in any systematic way

"Far more academic writing is produced for assessment purposes than for publication purposes, but because of the lack of a suitable corpus, research into the generic features of published academic writing vastly outweighs research into the generic features of assessed student writing" (Nesi et al. 2004: 440)

Michigan Corpus of Upper–level Student Papers
1. Introduction to MICUSP: Composition and markup

- Over 800 A–graded papers; around **2.3 million words** (May 2009 pre–release version)
- Papers collected from **16 disciplines** across 4 academic divisions (Humanities & Arts; Social Sciences; Biological & Health Sciences; Physical Sciences)
- Students at **4 levels** of study (senior undergraduates; 1st, 2nd, 3rd year graduates)
- Native and non–native speaker contributions
- Freely accessible online by the end of 2009

**Michigan Corpus of Upper–level Student Papers**
1. Introduction to MICUSP: Composition and markup

Distribution of papers across disciplines

Michigan Corpus of Upper-level Student Papers
1. Introduction to MICUSP: Composition and markup

MICUSP markup

- Each paper is encoded in TEI–compliant XML
- Combination of automatic and manual coding
- File header incorporating the metadata collected during paper submission
- **Structural divisions** (headings, sections, paragraphs) of the original paper maintained
- **Sentence tokenization**
- Encoding of textual features like quotations, emphasis, bullets
2. Positional variation of phrase–frames in MICUSP

2.1 Method

- Extracted n–grams and p–frames (spans: 3, 4, 5; floor=1) using \textit{kfNgram} (Fletcher 2002–2007)
- P–frames reduce n–gram lists in a motivated way
  - remove topic–specific items while highlighting discourse items
  - are more suited for the study of intra–textual variation
- Restricted definition of p–frames used here: only items with an internal variable slot (Römer, forthc.), not *BCD/ABC* type (Biber 2009)
- Different degrees of variability (VPR)
2. Positional variation of phrase–frames in MICUSP

2.1 Method – High–frequency p–frames

<table>
<thead>
<tr>
<th>Span 3</th>
<th>Span 4</th>
<th>Span 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>the * of the</td>
<td>the * of the</td>
<td>in the * of the</td>
</tr>
<tr>
<td>to * the</td>
<td>in the * of</td>
<td>of the * of the</td>
</tr>
<tr>
<td>the * and of</td>
<td>of the * of</td>
<td>to the * of the</td>
</tr>
<tr>
<td>a * of</td>
<td>to the * of</td>
<td>on the * of the</td>
</tr>
<tr>
<td>of * and the</td>
<td>and the * of</td>
<td>at the * of the</td>
</tr>
<tr>
<td>the * to the</td>
<td>the * of a</td>
<td>and the * of the</td>
</tr>
<tr>
<td>the * that the</td>
<td>on the * of</td>
<td>in order to the</td>
</tr>
<tr>
<td>the * in</td>
<td>of the * and</td>
<td>with the * and</td>
</tr>
<tr>
<td>and * the</td>
<td>for the * of</td>
<td>the * and the</td>
</tr>
<tr>
<td>the * is</td>
<td>the * and the</td>
<td>for the * of the</td>
</tr>
</tbody>
</table>

Michigan Corpus of Upper–level Student Papers
2. Positional variation of phrase–frames in MICUSP

2.1 Method – Recording textual position

- “textual colligation” (Hoey 2005)
- XML markup of document structure allows at least three kinds of position to be recorded:
  1. Position of 1st word of the item within its sentence
  2. Position of item’s sentence within the paragraph
  3. Position of the paragraph containing the item within the text

- Creation of a p-frame/n-gram and positional variation database from which distributions (for positions 1, 2, 3) can be retrieved

Michigan Corpus of Upper–level Student Papers
2. Positional variation of phrase-frames in MICUSP

2.1 Method – Position of phrase within sentence

- For each instance of p-frame in MICUSP, identify position of first word within its sentence
- Divide sentence into 3 parts: first and last 25%, remaining middle section (50%)

<table>
<thead>
<tr>
<th>Pos 1</th>
<th>Pos 2</th>
<th>Pos 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Michigan Corpus of Upper-level Student Papers
2. Positional variation of phrase–frames in MICUSP

2.1 Method – Position of sentence within paragraph

- For each instance of p-frame in MICUSP, identify location of containing sentence within its paragraph

- Divide paragraph into 3 parts: first (Pos 1) and last sentence of paragraph (Pos 3), sentences that are not first or last (Pos 2)
2. Positional variation of phrase–frames in MICUSP

2.1 Method – Position of paragraph within text

- For each instance of p-frame in MICUSP, identify location of containing paragraph within its text/paper

- Divide paragraph into 3 parts: Paragraph is within first 15% of paper (Pos 1), paragraph is part of mid–70% of paper (Pos 2), paragraph is within final 15% of paper (Pos 3)

<table>
<thead>
<tr>
<th>Pos 1</th>
<th>Pos 2</th>
<th>Pos 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *the * that

VPR: 33.49%

Variant/P-frame Ratio (Römer, forthc.)

- Avoidance of sentence-final position
- Even distribution across paragraphs
- Slight text-final preference

χ² ✓ p<0.001

Michigan Corpus of Upper-level Student Papers
2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *it is * that

VPR: 19.05%

- Strong sentence-initial preference
- Even distribution across paragraphs
- Preference for text-final position

\[ \chi^2 \]  
\[ p < 0.001 \]

**Michigan Corpus of Upper-level Student Papers**
2. Positional variation of phrase–frames in MICUSP

2.2 Select results: *it would be *to

-Strong preference for sentence–initial, and text–final positions

\[ \chi^2 \quad > \quad p < 0.001 \]

Michigan Corpus of Upper–level Student Papers
2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *as in figure*

VPR: 21.57%

- Strong preference for sentence-final position
- Mild paragraph- and text-final preference

\[ \chi^2 \checkmark \]
\[ p < 0.001 \]

Michigan Corpus of Upper-level Student Papers
2. Positional variation of phrase–frames in MICUSP

2.2 Select results: *in order to* *the*

- Strong preference for sentence– and paragraph– initial positions
- Mild text–medial preference
- Avoids sentence–final position

VPR: 61.64%

\[ \chi^2 \quad \checkmark \quad \checkmark \]

\[ p < 0.001 \]

Michigan Corpus of Upper–level Student Papers
2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *are likely to*

VPR: 7.63%

- Avoids sentence-final and favors sentence-medial/initial positions
- Strongly prefers paragraph-medial/final positions
- Mild text-medial preference

\[\chi^2 \checkmark \checkmark \]
\[p < 0.001\]

Michigan Corpus of Upper-level Student Papers
3. Conclusion

- We have introduced **MICUSP** as a new resource for the study of proficient student academic writing
- Use of p–frames as effective way to generalize phraseological items in advanced student writing
- Explored method to investigate textual distribution of p–frames

**Pedagogical implications**

- Important for novice writers to identify commonly used phrases in discourse of a discipline
- Important for EAP teachers and novice academic writers to know which items/phrases to use where in a text
3. Conclusion

Future avenues

- Look at the textual distribution of a larger number of items and use more rigorous statistical methods.
- Analyze p-frame internal variation (semantic grouping of variants).
- Study frequent items in context to examine their discourse functions.
- Carry out comparisons with expert/published writing (research articles from different disciplines) and with comparable sets of learner academic writing (availability issues).

Michigan Corpus of Upper-level Student Papers
Thank you!

Watch out for MICUSP coming soon!

Ute Römer & Matthew Brook O‘Donnell
uroemer@umich.edu, mbod@umich.edu

www.elicorpora.info
References


