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

Michigan Corpus Linguistics

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)

MICUSP

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- Gap: advanced but unpublished academic writing by graduate-level university students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 Difficulty of accessing unpublished academic writing--especially in any systematic way

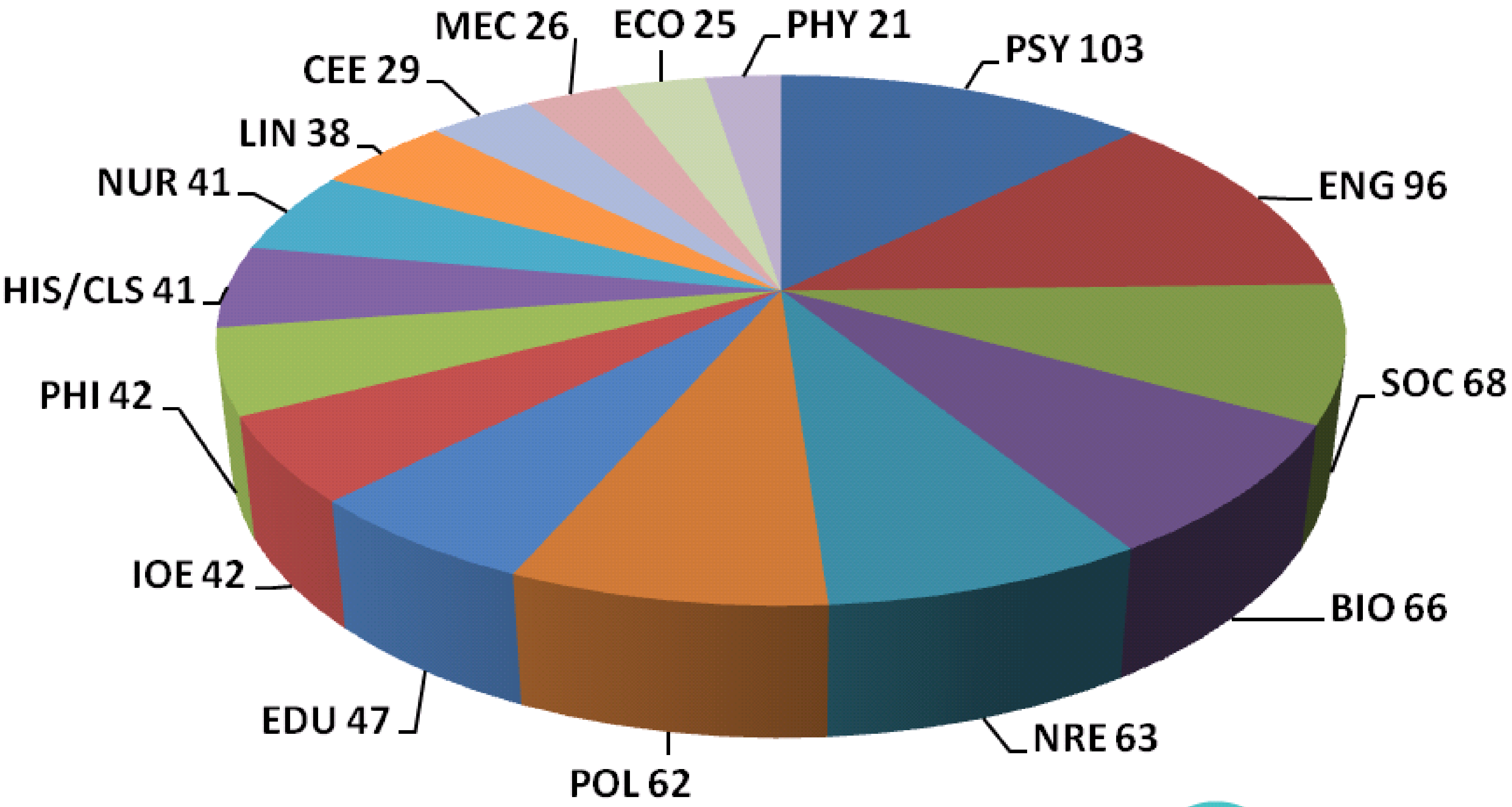
> "Far more academic writing is produced for assessment purposes than for publication purposes, but because of the lack of a suitable corpus, research into the generic features of published academic writing vastly outweighs research into the generic features of assessed student writing" (Nesi et al.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

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
- More info: flyer and http://micusp.elicorpora.info

1. Introduction to MICUSP: Composition and markup

Distribution of papers across disciplines



Michigan Corpus of Upper-level Student Papers

MICUSP

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 *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

Span 3	Span 4	Span 5
*the * of*	*the * of the*	*in the * of the*
*to * the*	*in the * of*	*of the * of the*
*the * and*	*of the * of*	*to the * of the*
*a * of*	*to the * of*	*on the * of the*
*of * and*	*and the * of*	*at the * of the*
*the * to*	*the * of a*	*and the * of the*
*the * that*	*on the * of*	*in order to * the*
*the * in*	*of the * and*	*with the * and the*
*and * the*	*for the * of*	*of the * and the*
*the * is*	*the * and the*	*for the * of the*

2. Positional variation of phrase-frames in MICUSP

2.1 Method – Recording textual position

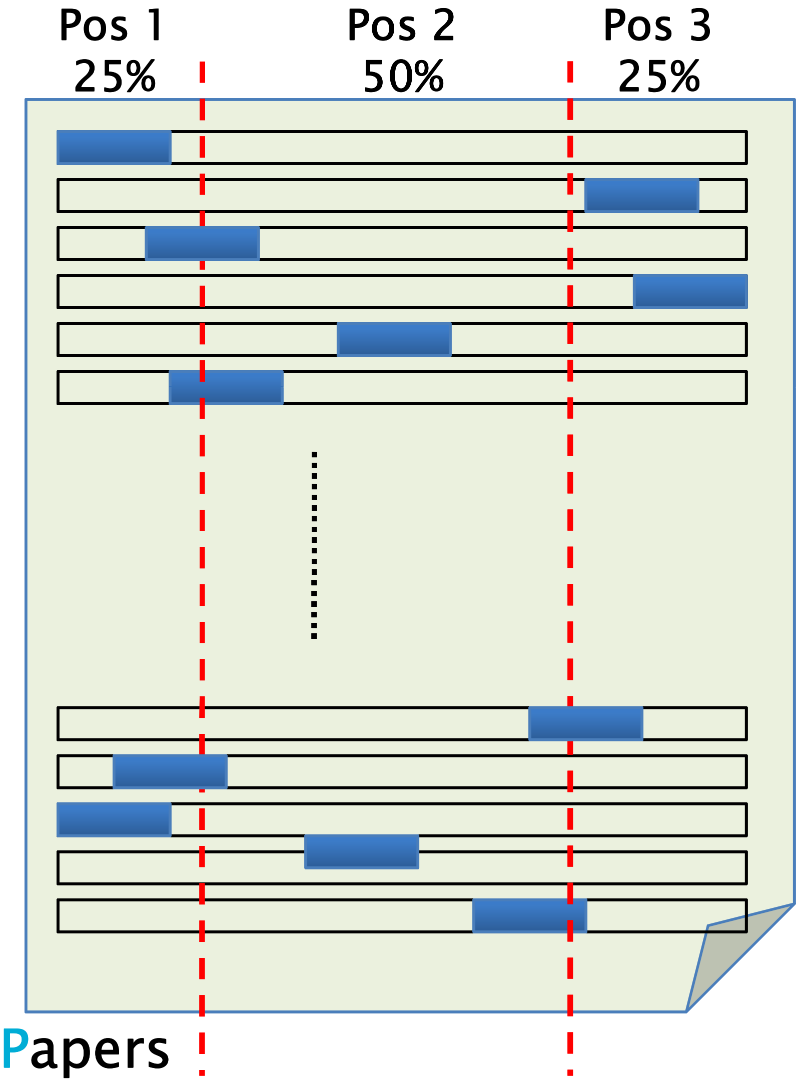
- “textual colligation” (Hoey 2005)
- XML markup of document structure allows at least three kinds of position to be recorded:
  1. Position of 1st word of the item within its sentence
  2. Position of item’s sentence within the paragraph
  3. Position of the paragraph containing the item within the text
- Creation of a p-frame/n-gram and positional variation database from which distributions (for positions 1, 2, 3) can be retrieved

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%)

Pos 1	Pos 2	Pos 3
5	4	2

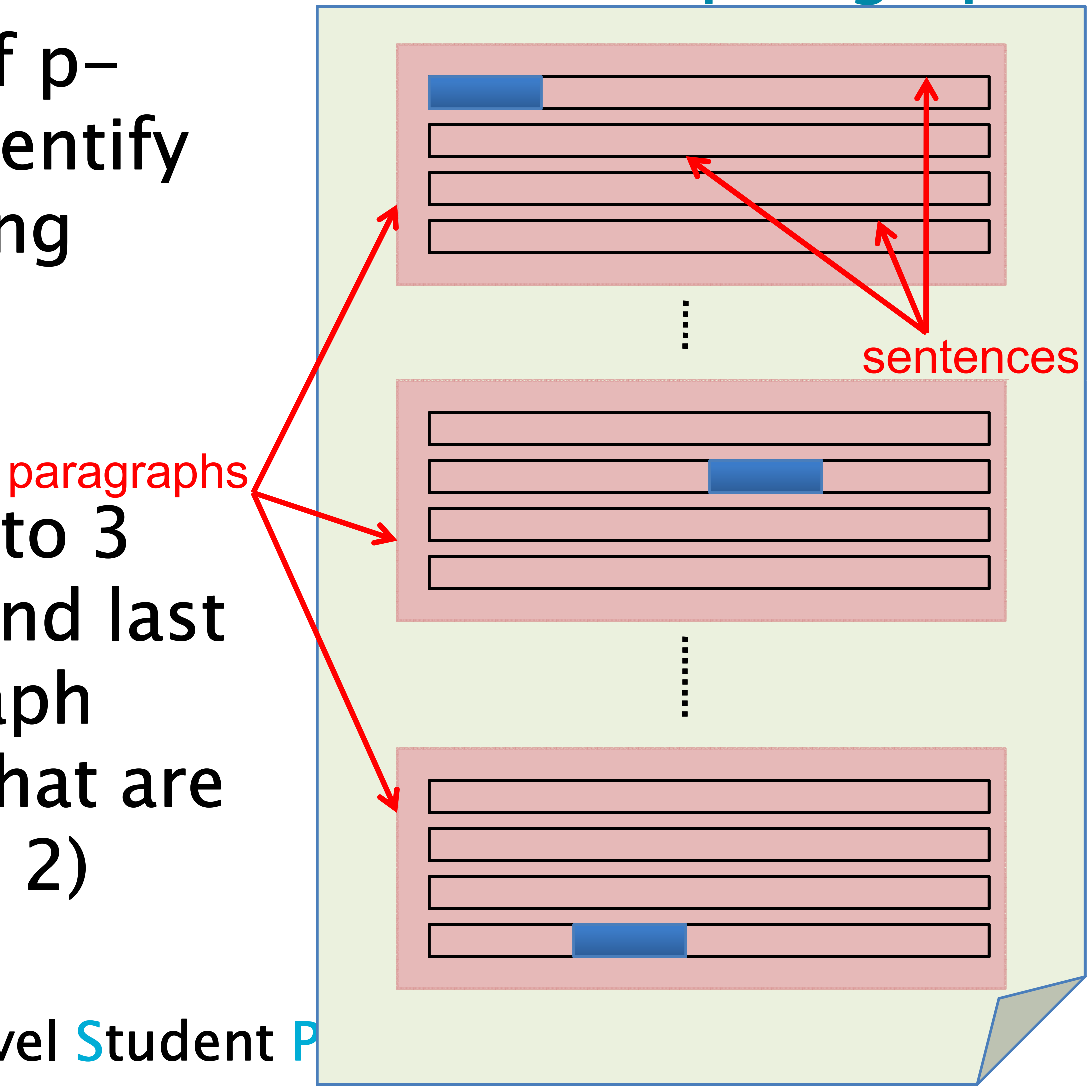


Michigan Corpus of Upper-level Student Papers

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)

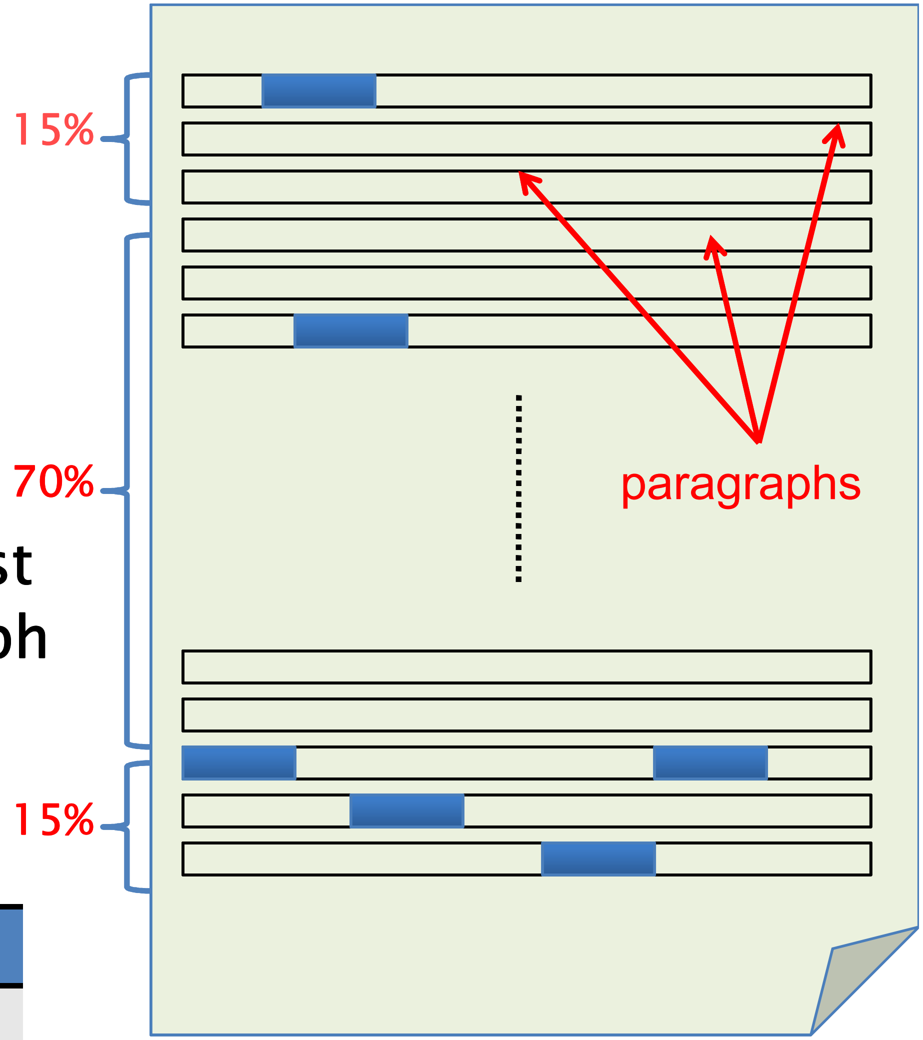


2. Positional variation of phrase-frames in MICUSP

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)

Pos 1	Pos 2	Pos 3
1	1	4

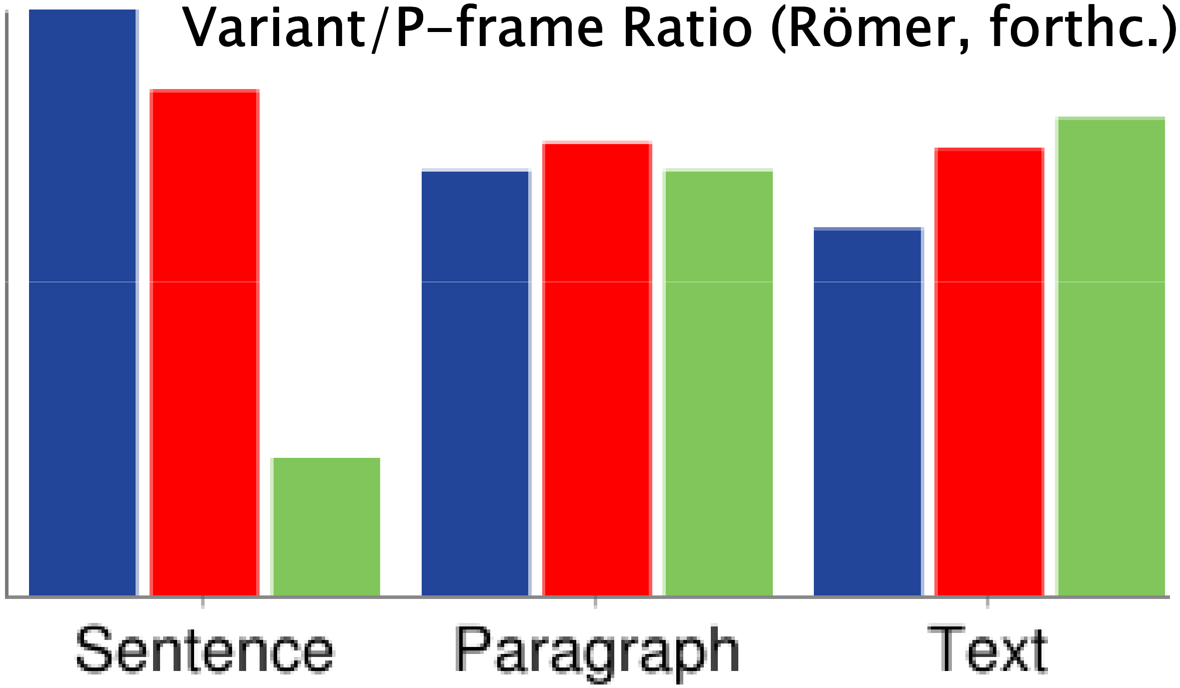


2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *the * that*

VPR: 33.49%

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



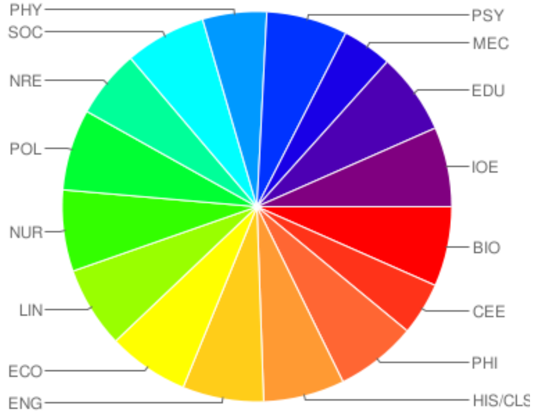
χ^2 ✓ ✓

p<0.001

-Avoidance of sentence-final position

-Even distribution across paragraphs

-Slight text-final preference



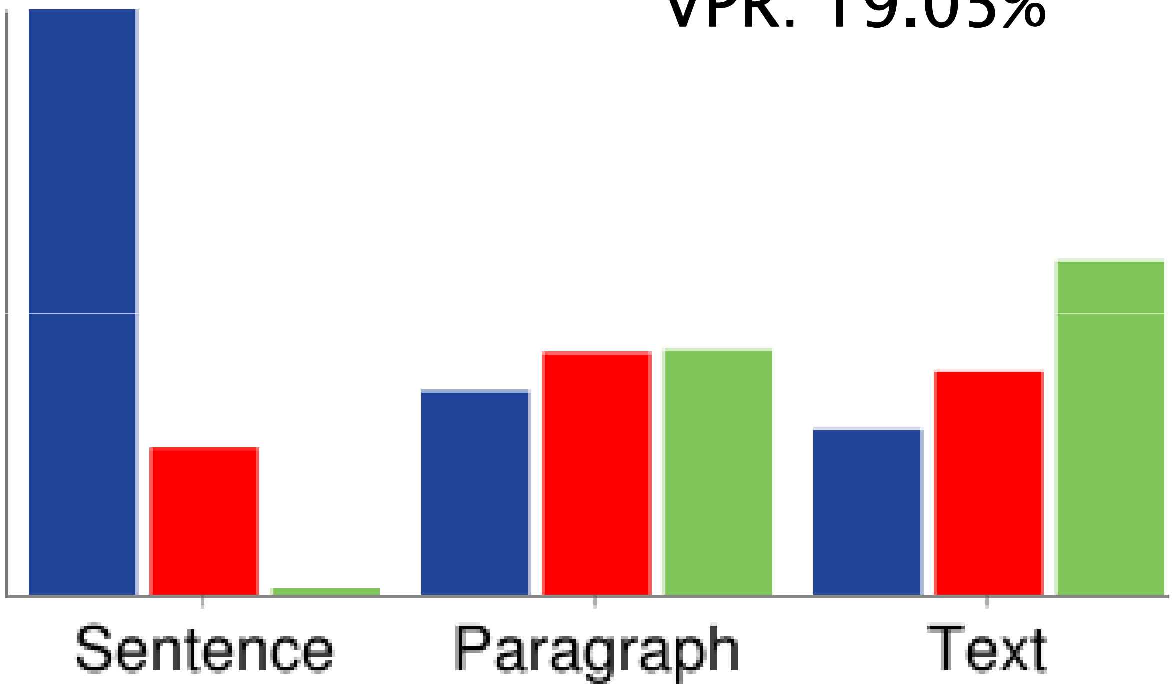
MICUSP

Michigan **C**orpus of **U**pper-level **S**tudent **P**apers

2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *it is * that*

VPR: 19.05%



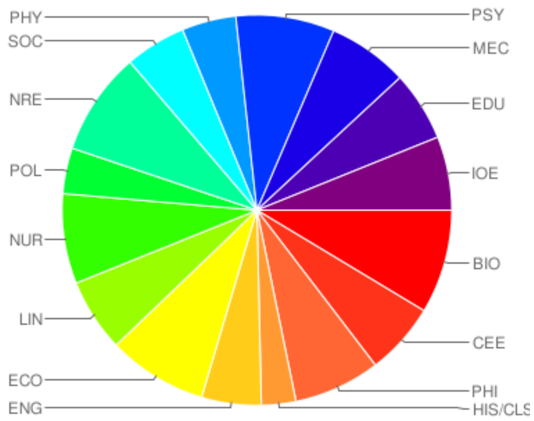
χ^2 ✓ ✓

$p<0.001$

-Strong sentence-initial preference

-Even distribution across paragraphs

-Preference for text-final position



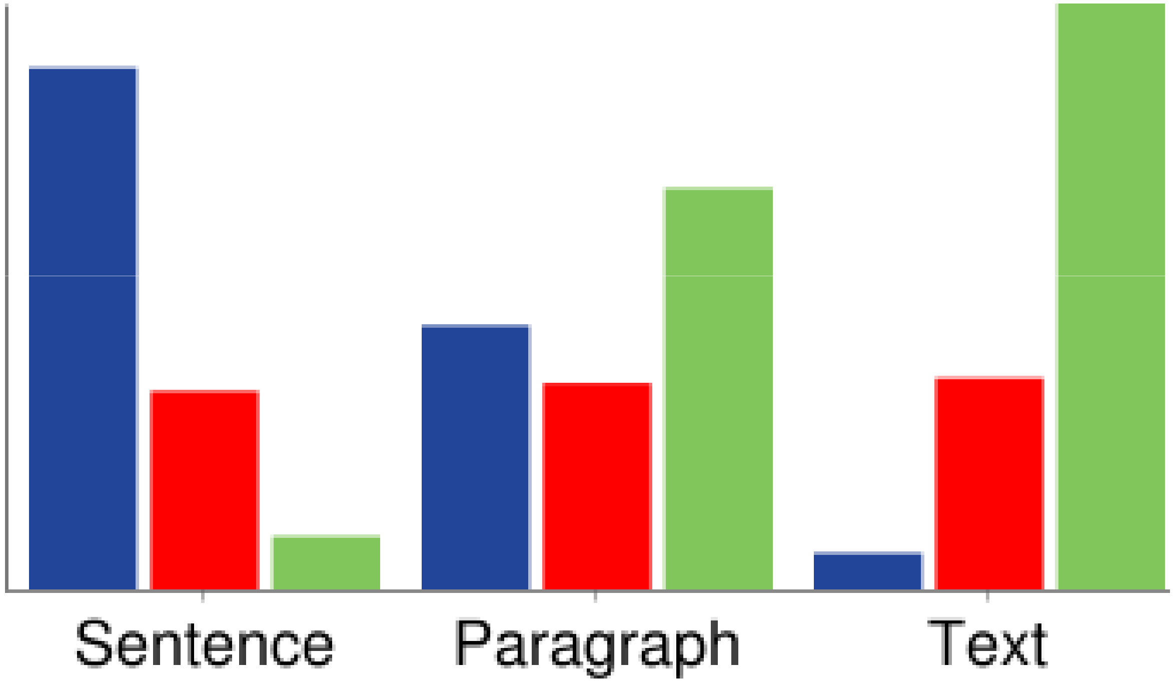
Michigan Corpus of Upper-level Student Papers

MICUSP

2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *it would be * to*

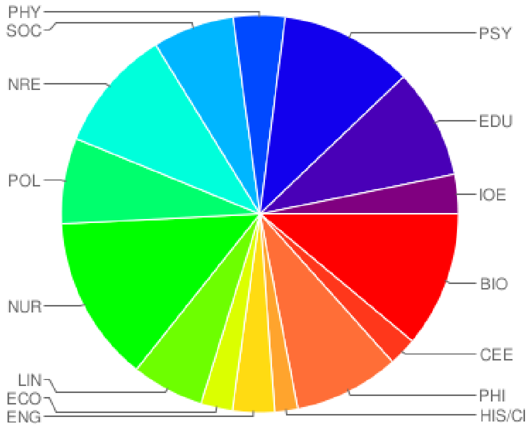
VPR: 37.25%



χ^2 ✓ ✓

$p<0.001$

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



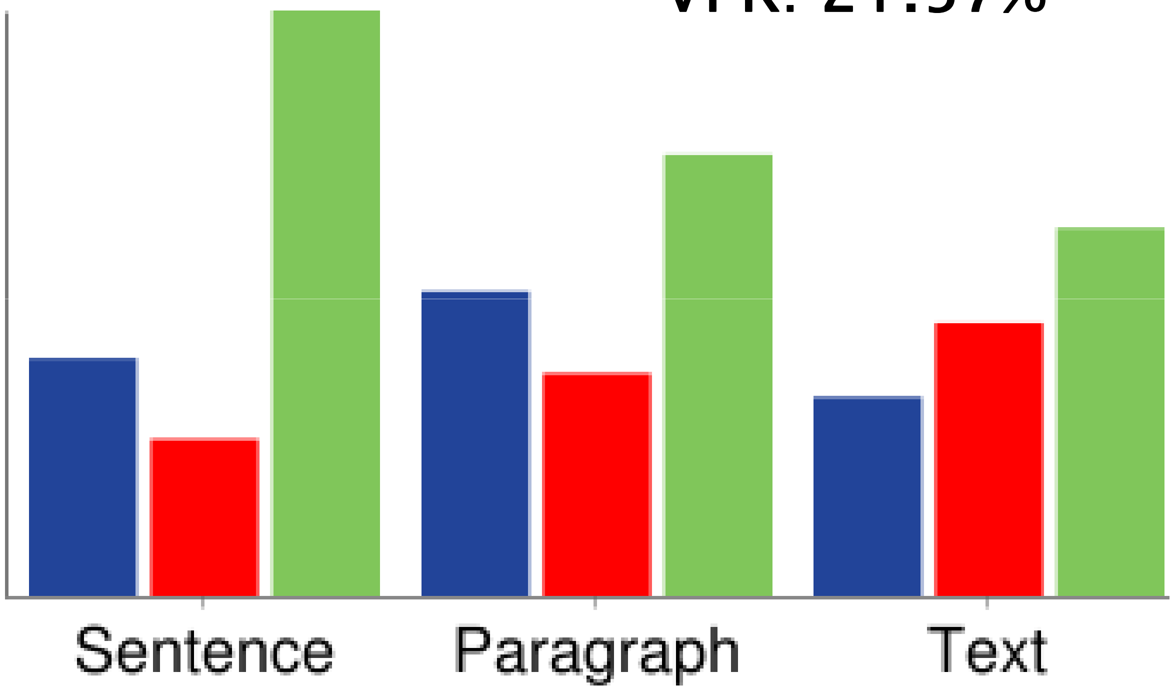
Michigan Corpus of Upper-level Student Papers

MICUSP

2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *as * in figure*

VPR: 21.57%

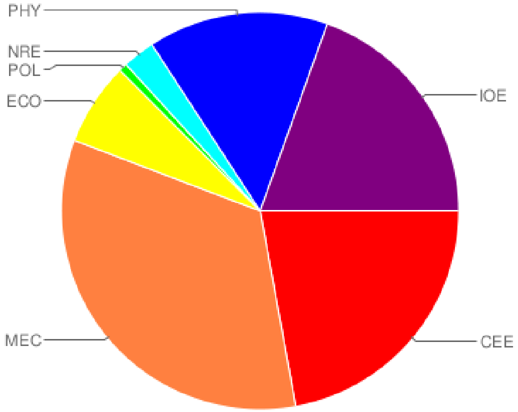


χ^2 ✓

$p<0.001$

-Strong preference for sentence-final position

-Mild paragraph- and text-final preference



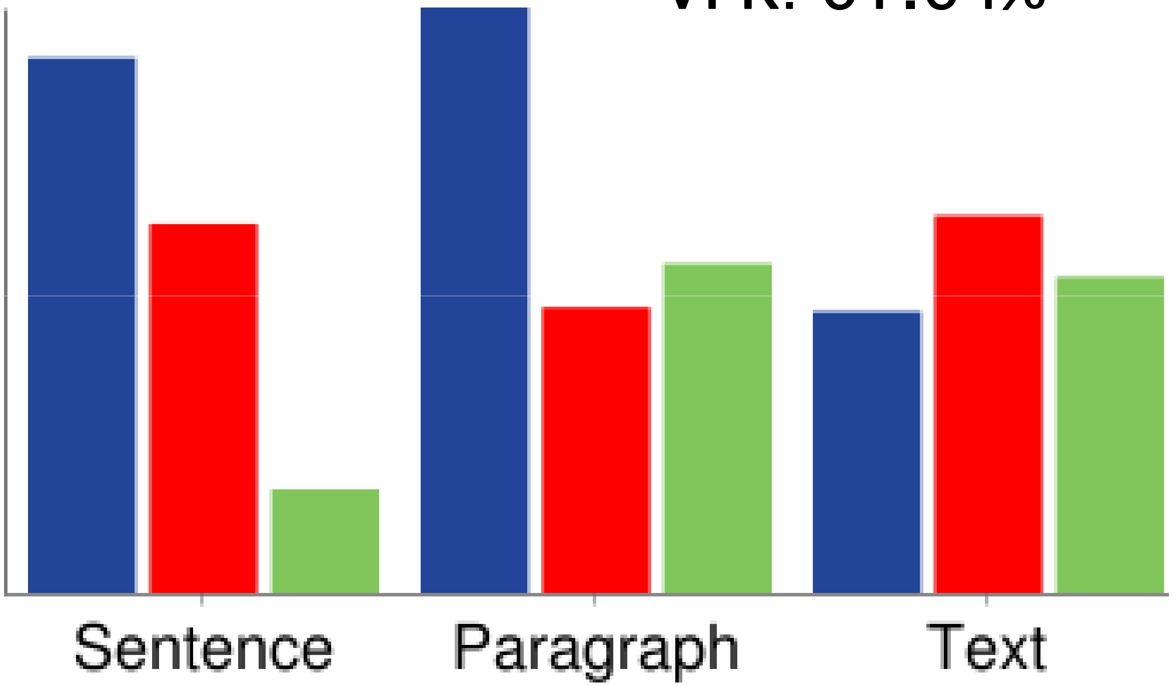
Michigan Corpus of Upper-level Student Papers

MICUSP

2. Positional variation of phrase-frames in MICUSP

2.2 Select results: *in order to * the*

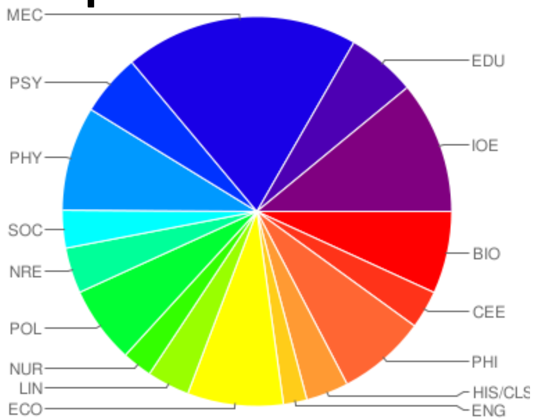
VPR: 61.64%



χ^2 ✓ ✓

p<0.001

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



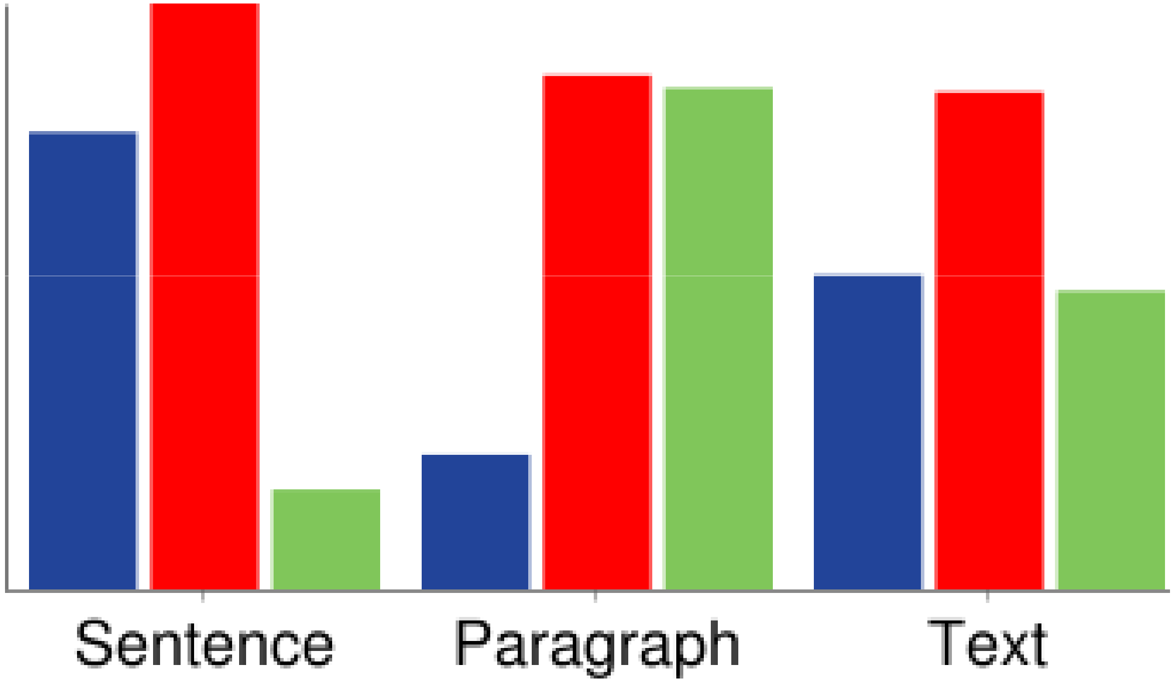
MICUSP

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%



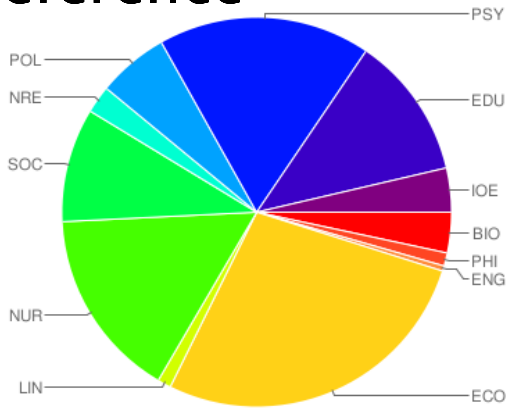
χ^2 ✓ ✓

$p<0.001$

-Avoids sentence-final and favors sentence-medial/ initial positions

-Strongly prefers paragraph-medial/ final positions

-Mild text-medial preference



Michigan Corpus of Upper-level Student Papers

MICUSP

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 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)

Thank you!

Watch out for MICUSP coming soon!

Ute Römer & Matthew Brook O‘Donnell

uroemer@umich.edu, mbod@umich.edu

www.elicorpora.info

Michigan Corpus Linguistics

References

Ädel, A. and U. Römer. (In preparation). Research on proficient academic writing across disciplines and levels: Introducing MICUSP.
Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3): 275-311.
Fletcher, W. H. (2002-2007). *KfNgram*. Annapolis, MD: USNA.
Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
Hoey, M. and M. B. O'Donnell (2008). 'The Beginning of Something Important? Corpus Evidence on the Text Beginnings of Hard News Stories'. In: Lewandowska-Tomaszczyk, B. (ed.), *Corpus Linguistics, Computer Tools, and Applications-State of the Art. PALC 2007*. Bern: Peter Lang.
Nesi, H., G. Sharpling and L. Ganobcsik-Williams. (2004). Student papers across the curriculum: Designing and developing a corpus of British student writing. *Computers and Composition* 21: 439-450.
O'Donnell, M. B. and U. Römer. (In preparation). From student hard drive to web corpus: The design, compilation, annotation and online distribution of MICUSP.
Römer, U. (Forthcoming). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1).
Römer, U. and S. Wulff. (Forthcoming). Applying corpus methods to writing research: Explorations of MICUSP. *Journal of Writing Research* (http://www.jowr.org).
Wulff, S. and U. Römer. (Forthcoming). Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora*.
Wulff, S., U. Römer and J. M. Swales. (In preparation). Attended/unattended *this* in academic writing: Qualitative and quantitative perspectives. *Corpus Linguistics and Linguistic Theory*.