

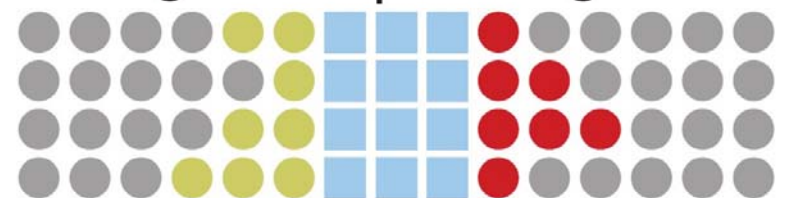
# Analyzing the disciplinary and textual distribution of phraseological items in a new corpus of proficient student writing

Matthew Brook O'Donnell & Ute Römer

AAAL 2010, Atlanta – 7 March 2010

[www.elicorpora.info](http://www.elicorpora.info)

Michigan Corpus Linguistics



# Presentation outline

1. Introduction to MICUSP (Michigan Corpus of Upper-level Student Papers): Composition and markup
2. Phraseological variation across MICUSP disciplines
3. Positional variation of select phraseological items
4. 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- Gap: advanced but unpublished academic writing by graduate-level university students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 Difficulty of accessing unpublished academic writing--especially in any systematic way

# 1. Introduction to MICUSP: Composition and markup

- 829 A–graded papers; around **2.6 million words**
- 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 using **MICUSP Simple**
- **More info: flyer and <http://micusp.elicorpora.info>**

# 1. Introduction to MICUSP: Composition and markup

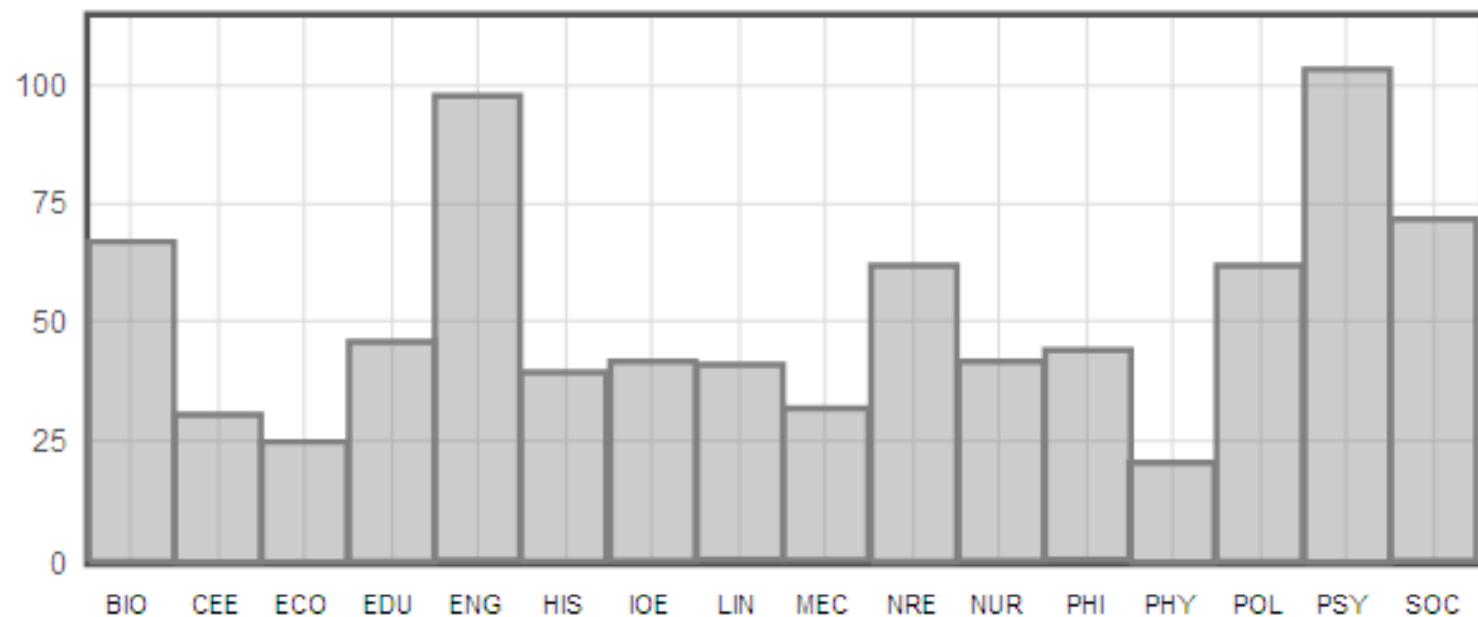
## Distribution of papers across disciplines

 **MICUSP Simple** <sup>BETA</sup>  
Michigan Corpus of Upper-Level Student Papers

### DISTRIBUTION ACROSS DISCIPLINES

CLICK TO SELECT

Clear selection



Michigan Corpus of Upper-level Student Papers



# 1. Introduction to MICUSP: Composition and markup

## MICUSP markup

- Each paper is encoded in TEI-compliant XML
- File header incorporating the metadata collected during paper submission
- Encoding of textual features like quotations, emphasis, bullets
- **Structural divisions** (headings, sections, paragraphs) of the original paper maintained
- **Sentence tokenization**

## 2. Phraseological variation across MICUSP disciplines

### The importance of phraseology in linguistics

- "the normal carrier of meaning is the phrase" (Sinclair 2005)
- Phraseology: the heart of language (Ellis 2008)
- In academic writing: important to know genre- and discipline-specific phrases
- Aim: construct a phraseological profile of student academic writing across disciplines; focus on **identification** and **textual distribution** of items

## 2. Phraseological variation across MICUSP disciplines

### Data and analytic procedure

- Phraseology measured through n-grams and p-frames
- N-grams: *on the one hand, on the other hand*
- P-frames (*on the \* hand*) reduce n-gram lists in a motivated way
  - remove topic-specific items while highlighting discourse items so are 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)

## 2. Phraseological variation across MICUSP disciplines

### Data and analytic procedure

- Extracted n-grams and p-frames (spans: 3, 4, 5) using *kfNgram* (Fletcher 2002–2007)
- N-gram/p-frame extractions based on:
  - MICUSP\_all
  - 4 MICUSP subsets: **BIO, ENG, MEC/IOE, PSY**
- Created **key-PI lists** for each MICUSP subset/span (MICUSP\_all as reference corpus)

## 2. Phraseological variation across MICUSP disciplines

### Selected results: high-frequency PIs in MICUSP

Span 3	Span 4	Span 5
<i>it would be</i> <i>the fact that</i> <i>as well as</i> <i>the * of</i> <i>to * the</i> <i>the * and</i> <i>the * to</i> <i>the * that</i>	<i>on the other hand</i> <i>at the same time</i> <i>as a result of</i> <i>the * of the</i> <i>in the * of</i> <i>it is * to</i> <i>at the * of</i> <i>it is * that</i>	<i>in the * of the</i> <i>of the * of the</i> <i>at the * of the</i> <i>in order to * the</i> <i>it is * that the</i> <i>for the * of the</i> <i>with the * and the</i> <i>on the * of the</i>

## 2. Phraseological variation across MICUSP disciplines

### Selected results: key-PIs across disciplines

- General finding: lots of topic-related PIs that only occur in the selected discipline, e.g.

**BIO:** *of \* species, the \* of cholera, body color \* wing venation*

**ENG:** *in \* novel, the vicar of wakefield, the \* of st albans*

**MEC/IOE:** *the zero dynamics, of \* vehicle, in the front end*

**PSY:** *the psychological well-being, in the \* game, of child and adolescent*

- Look at non-topic/discourse related items...

## 2. Phraseological variation across MICUSP disciplines

### Key-PIs within disciplines: BIOLOGY

Span 3	Span 4	Span 5
<i>stark et al et al 2000 total number of in * experiment the number of</i>	<i>stark et al 1966 the total number of</i>	<i>it has been * that</i>  – Use of sources – Bibliographic conventions – Empirical research

## 2. Phraseological variation across MICUSP levels and disciplines

### Key-PIs within disciplines: ENGLISH

- Almost all key-PIs topic-related
- Many pronoun clusters (e.g. *of her own, she is not, his \* and his*), probably from papers about novels/plays
- *in the \* of, the \* in which, at the end* – partly topic-related
- Negative key items: *due to the, based on the, more likely to, the number of*

## 2. Phraseological variation across MICUSP levels and disciplines

### Key-PIs within disciplines: ENGINEERING (MEC/IOE)

Span 3	Span 4	Span 5
<i>we * that shown in figure we determined the layout * the the * model to calculate the table * below</i>	<i>as * in figure shown in * 1 we were * to it was * that we were able to as shown in figure</i>	<i>of * project is to it was * that the in order to * the</i>  – Empirical research – Reporting findings – Focus on analytic procedure & results

## 2. Phraseological variation across MICUSP levels and disciplines

### Key-PIs within disciplines: PSYCHOLOGY

Span 3	Span 4	Span 5
<i>more likely to higher * of been shown that in * study may be more more likely to of the study et al 2002 in * research have found that</i>	<i>are * likely to the * effects of are more likely to</i>  <b>Negative key item:</b> <i>the * of the</i>	[none; all topic-related]  – Literature review/ use of sources – Hedging – Empirical research

### 3. Positional variation of select phraseological items

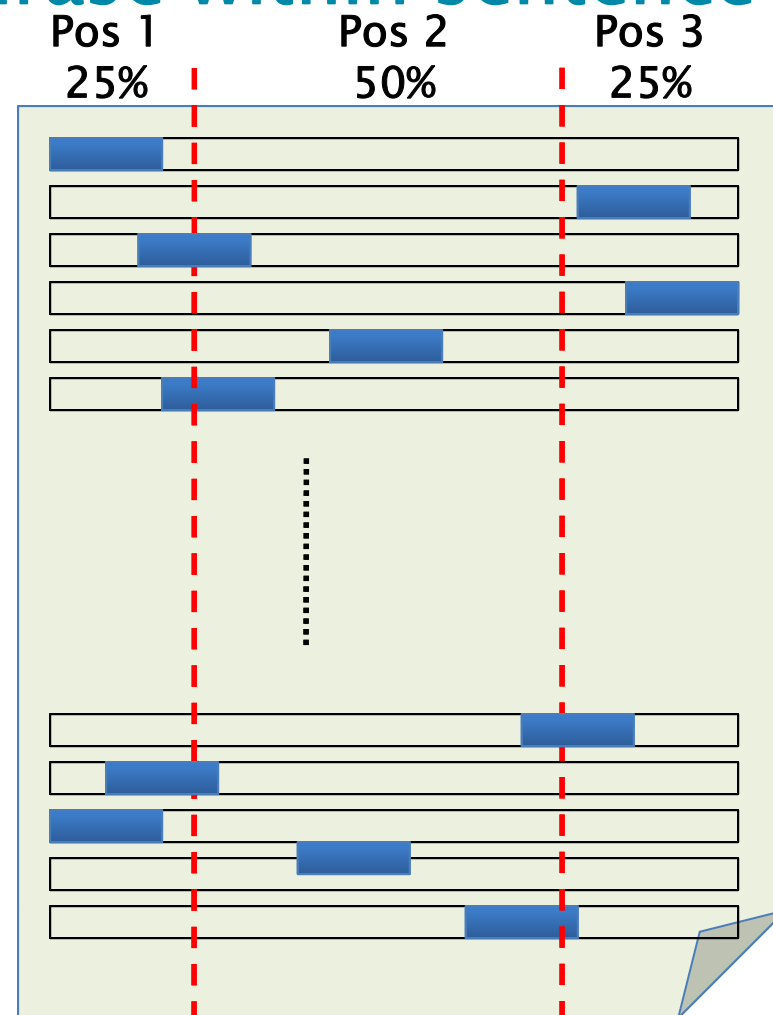
- Hoey (2005) suggests that words and phrases have **associations with particular textual positions**, e.g. the beginning or end of sentences, paragraphs and texts
  - Textual Priming Project (Hoey & O'Donnell 2008) confirmed this claim in the analysis of a large newspaper corpus (e.g. *according to a*)
- Here we apply similar analysis to PIs in MICUSP
- Method: Creation of a **p-frame/n-gram and positional variation database** from which a PI's sentence, paragraph and text positional profile can be retrieved

### 3. Positional variation of select phraseological items

#### 1. Position of (first word of) phrase within sentence

- For each instance of phraseological item 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



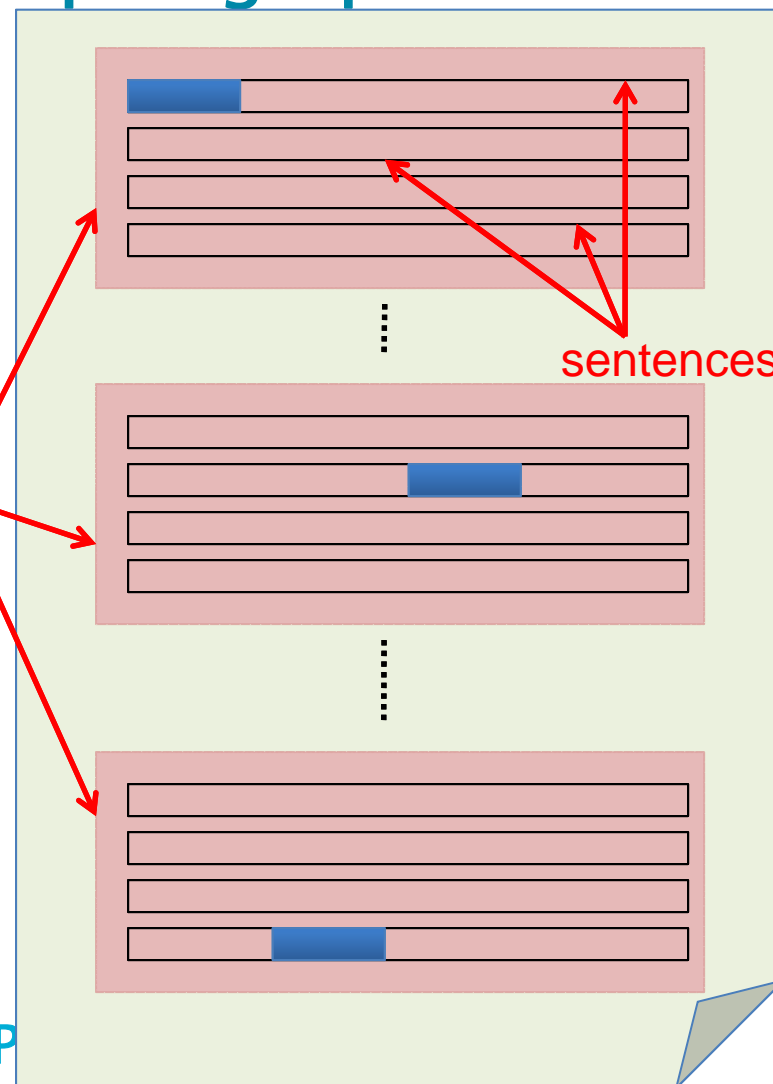
### 3. Positional variation of select phraseological items

#### 2. Position of sentence within paragraph

- For each instance of phraseological item 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)

paragraphs

sentences

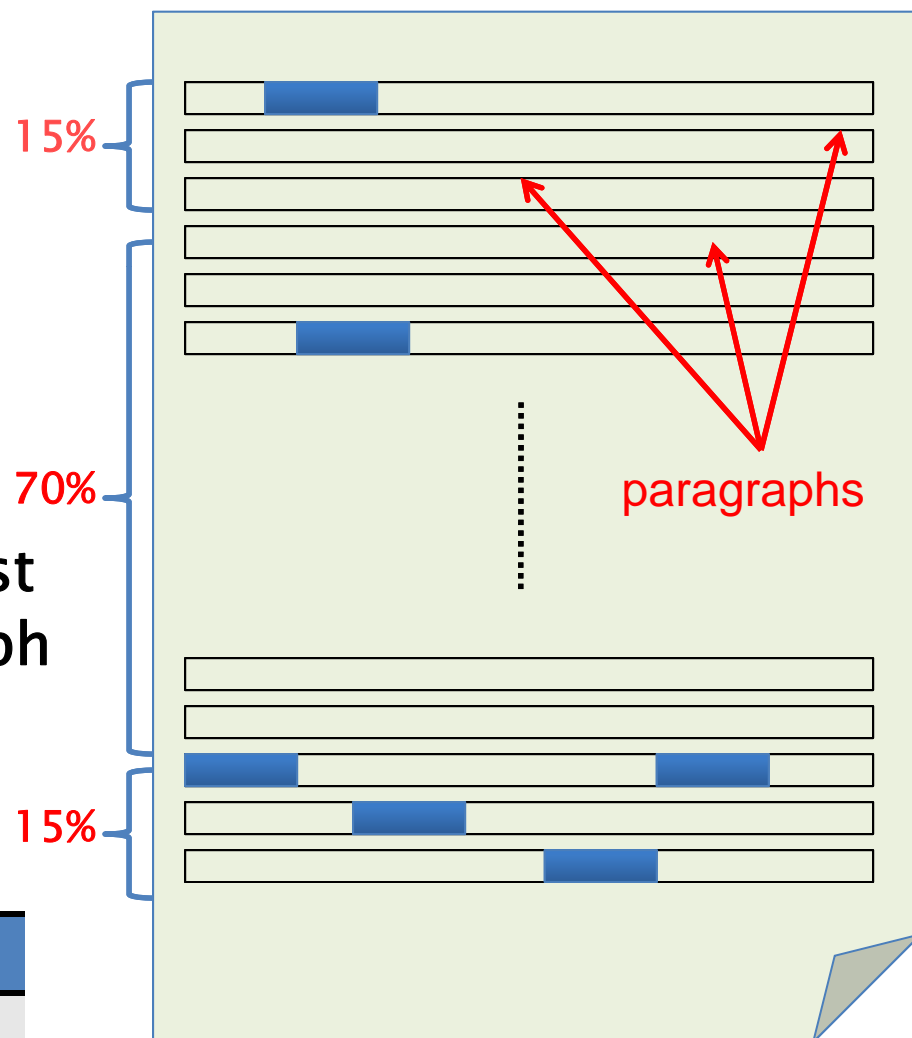


# 3. Positional variation of select phraseological items

## 3. Position of paragraph within text

- For each instance of phraseological item 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

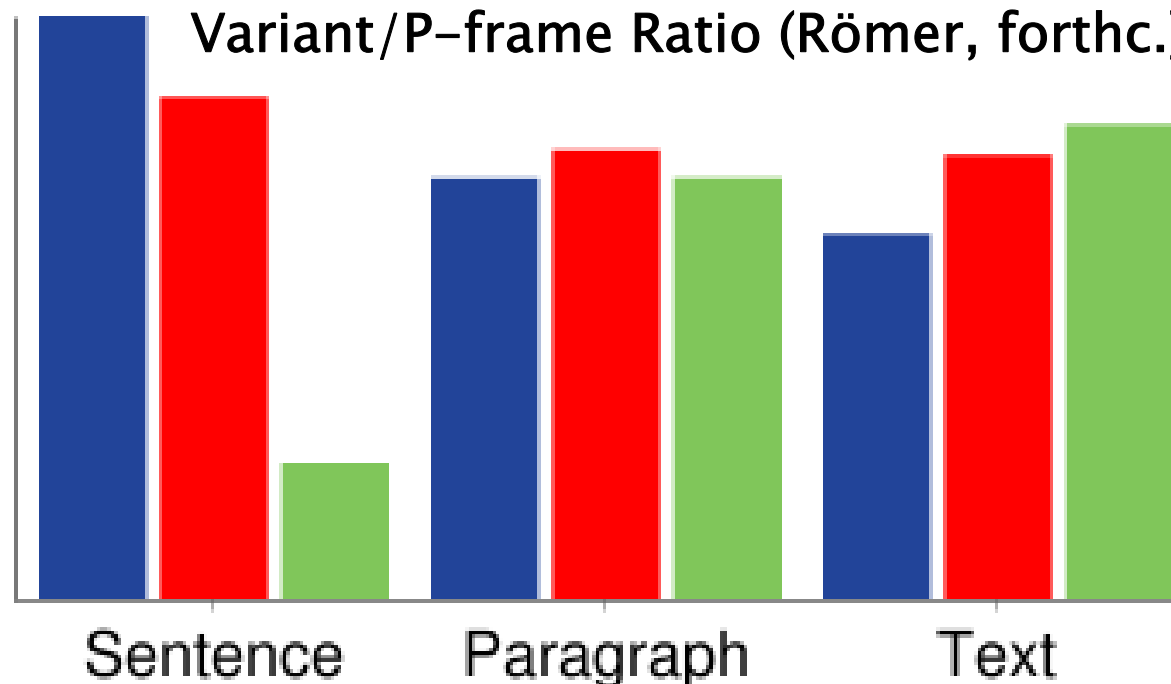


### 3. Positional variation of select phraseological items

Select results: *the \* that*

VPR: 33.49%

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

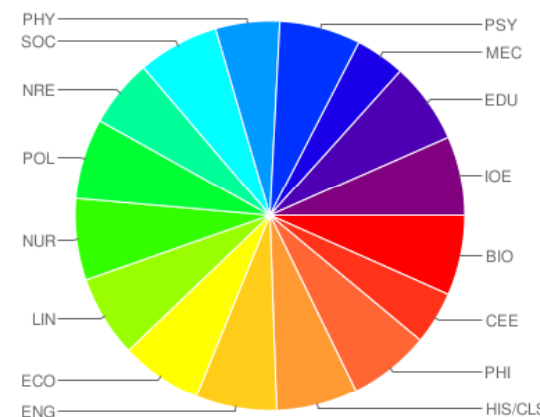


$\chi^2$  ✓  
p<0.001

-Avoidance of sentence-final position

-Even distribution across paragraphs

-Slight text-final preference

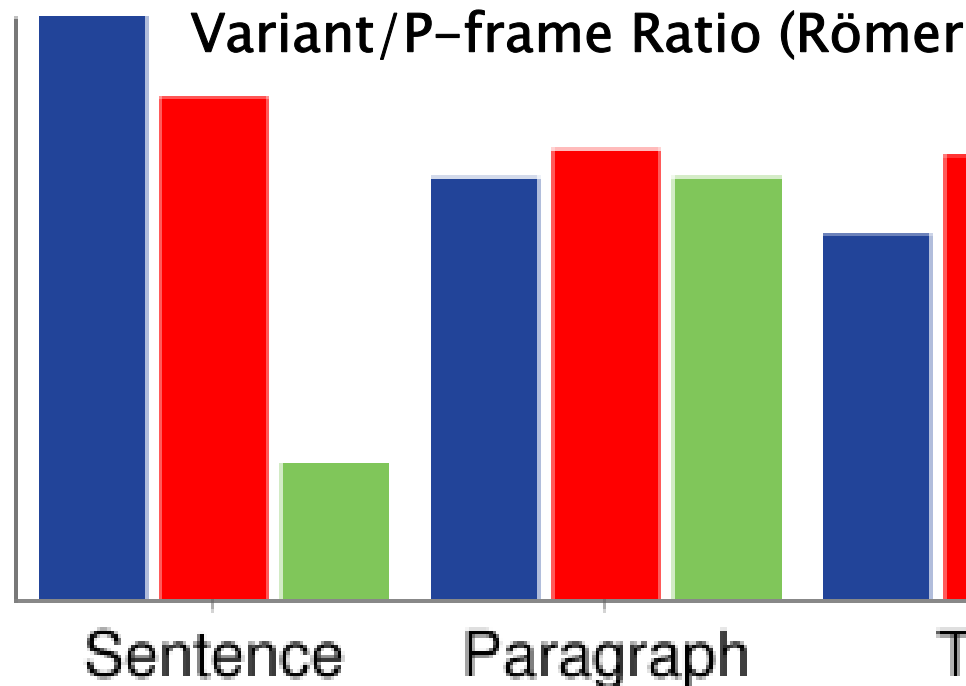


### 3. Positional variations of select phraseological items

Select results: *the \* that*

VPR: 33.49%

Variant/P-frame Ratio (Römer)



$\chi^2$  ✓  
p < 0.001

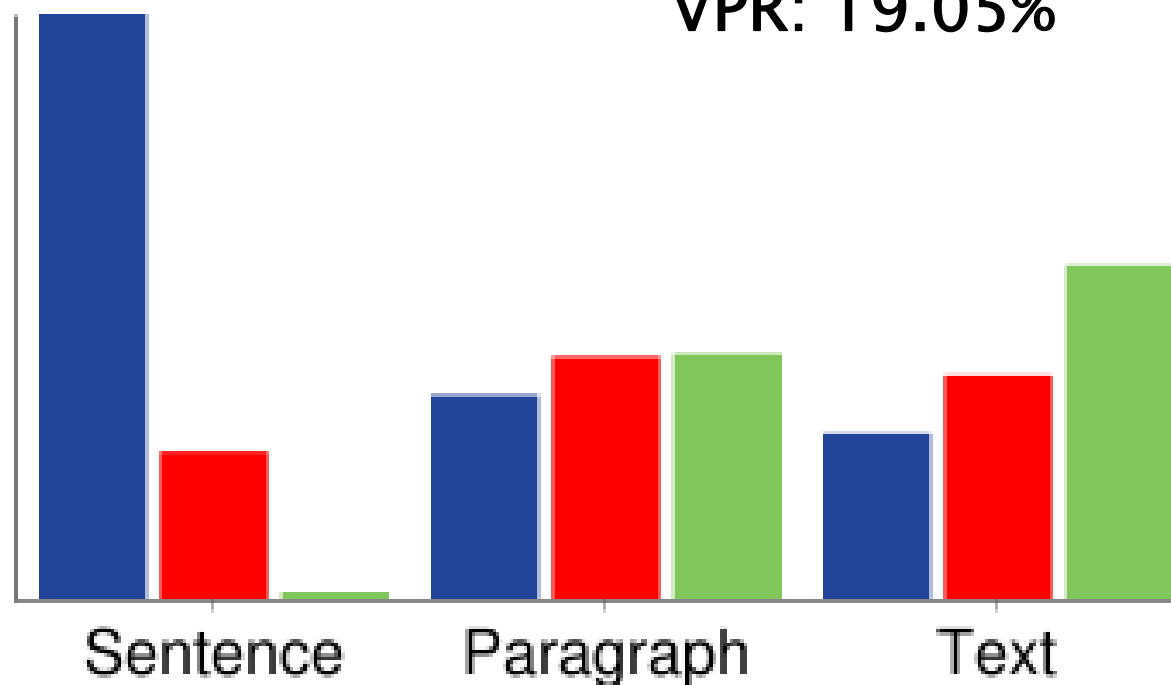
VPR: 33.49% - Tokens: 2905 - Variants: 973

fact	451
idea	134
way	64
assumption	51
notion	50
possibility	44
ways	42
belief	41
conclusion	38
sense	34
extent	28
claim	27
case	26
point	24
reader	23
role	23
factors	22

### 3. Positional variation of select phraseological items

Select results: *it is \* that*

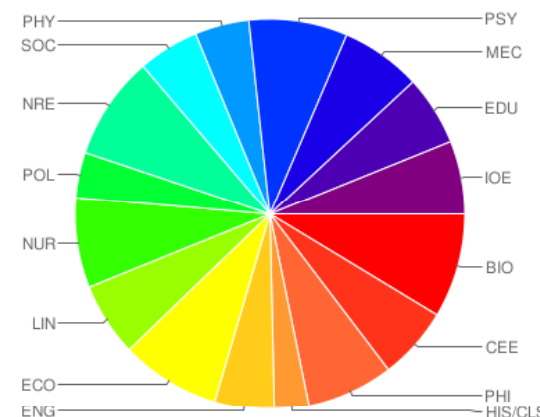
VPR: 19.05%



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

$\chi^2$  ✓  
p<0.001

✓



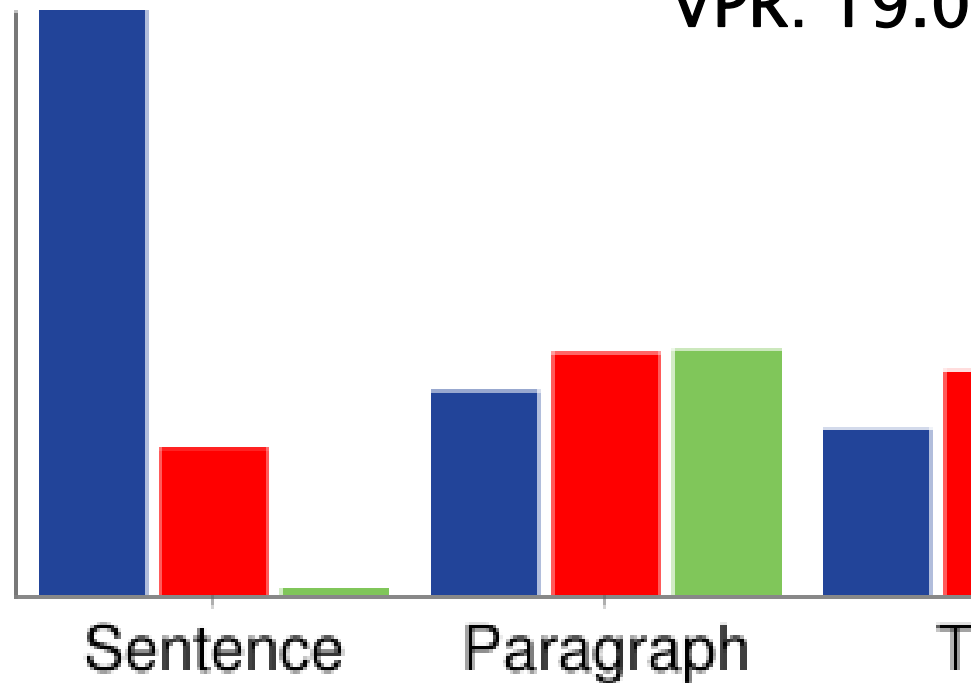
Michigan Corpus of Upper-level Student Papers



### 3. Positional variation of select phraseological items

Select results: *it is \* that*

VPR: 19.0



$\chi^2$  ✓  
p < 0.001

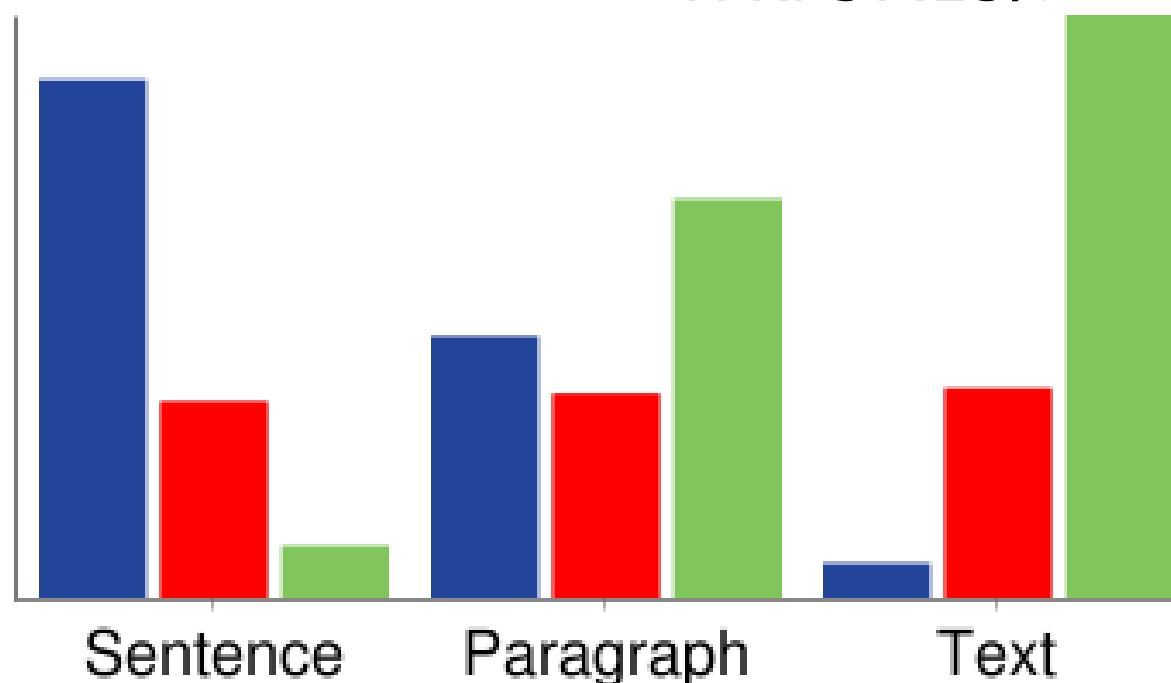
VPR: 19.05% - Tokens: 588 - Variants: 112

clear	90
possible	72
likely	36
important	27
true	22
unlikely	19
assumed	17
apparent	16
expected	13
obvious	13
interesting	11
estimated	11
evident	10
hypothesized	8
crucial	8
known	8
probable	7

### 3. Positional variation of select phraseological items

Select results: *it would be \* to*

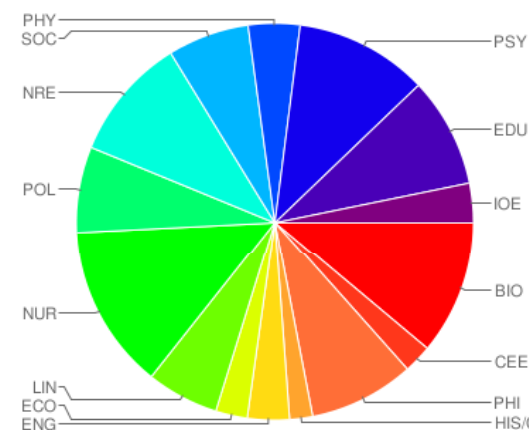
VPR: 37.25%



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

$\chi^2$  ✓  
p<0.001

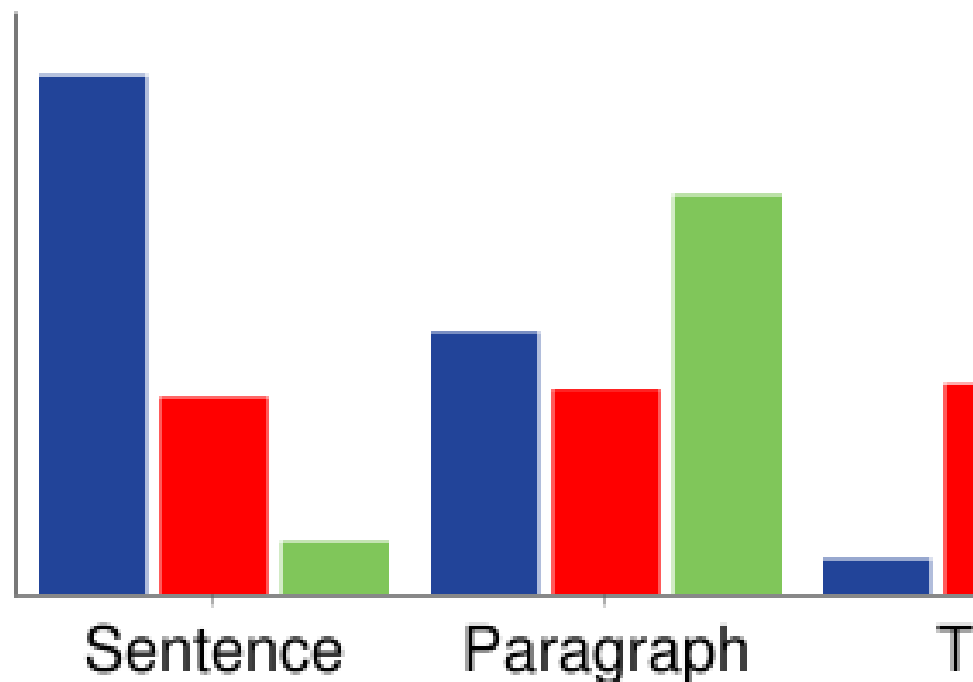
✓



### 3. Positional variation of select phraseological items

Select results: *it would be*

VPR: 37.2



$\chi^2$  ✓  
p < 0.001

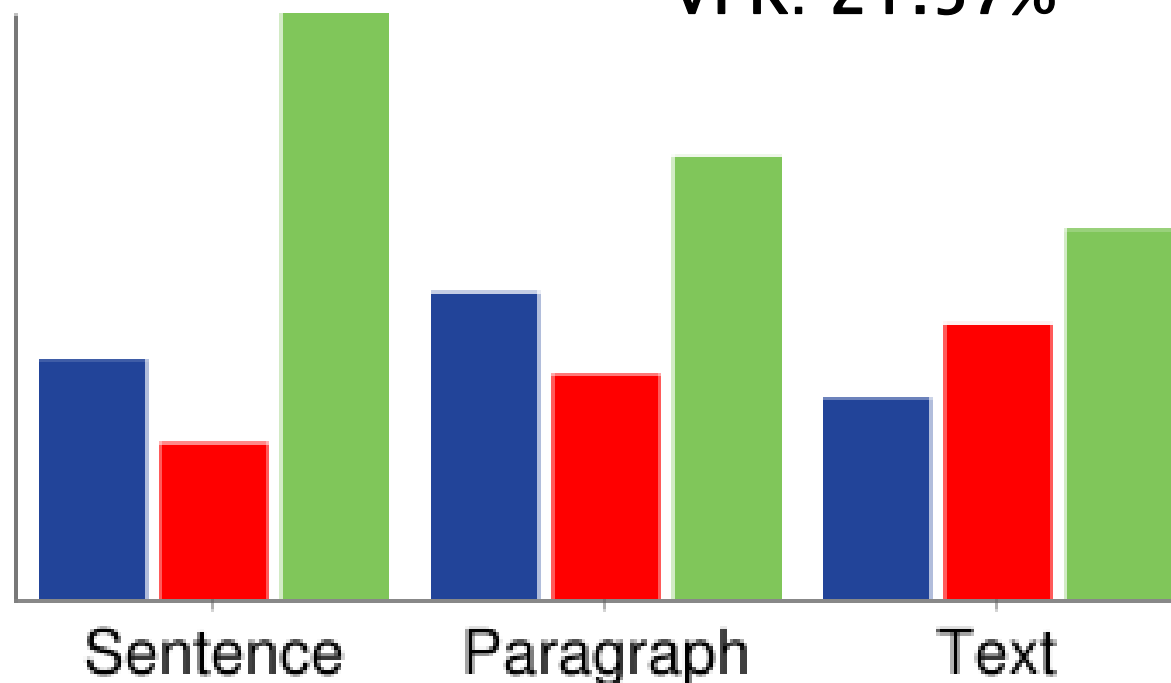
VPR: 37.25% - Tokens: 102 - Variants: 38

interesting	17
hard	9
difficult	9
useful	6
easy	4
important	4
helpful	4
worthwhile	3
unfair	3
easier	3
possible	3
false	2
appropriate	2
impossible	2
best	2
preferable	2

### 3. Positional variation of select phraseological items

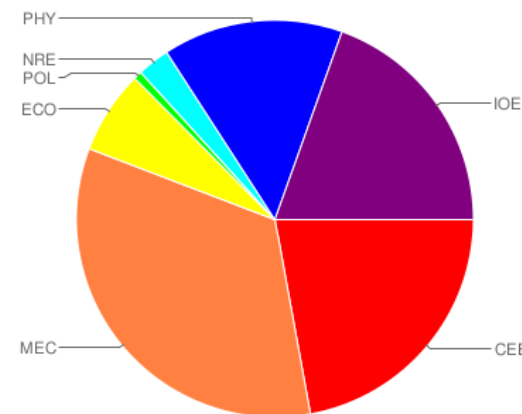
Select results: *as* \* *in figure*

VPR: 21.57%



–Strong preference for sentence–final position

–Mild paragraph– and text–final preference

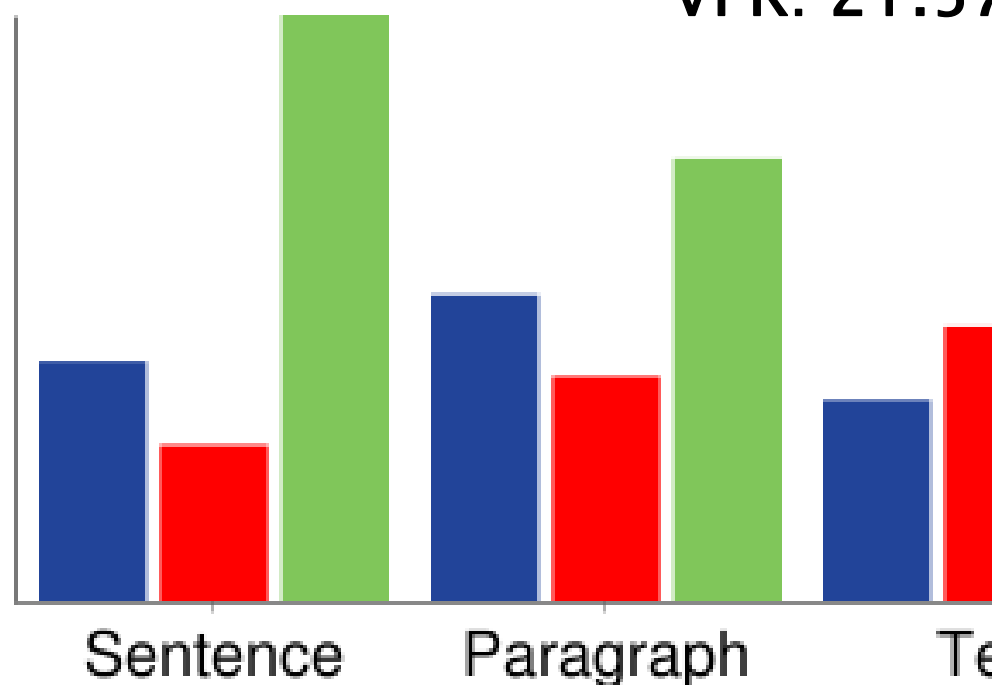


$\chi^2$  ✓  
p<0.001

### 3. Positional variation of select phraseological items

Select results: *as \* in figure*

VPR: 21.57%



VPR: 21.57% - Tokens: 51 - Variants: 11

shown	30
seen	9
described	3
illustrated	2
used	1
visualized	1
displayed	1
observed	1
depicted	1
x	1
y	1

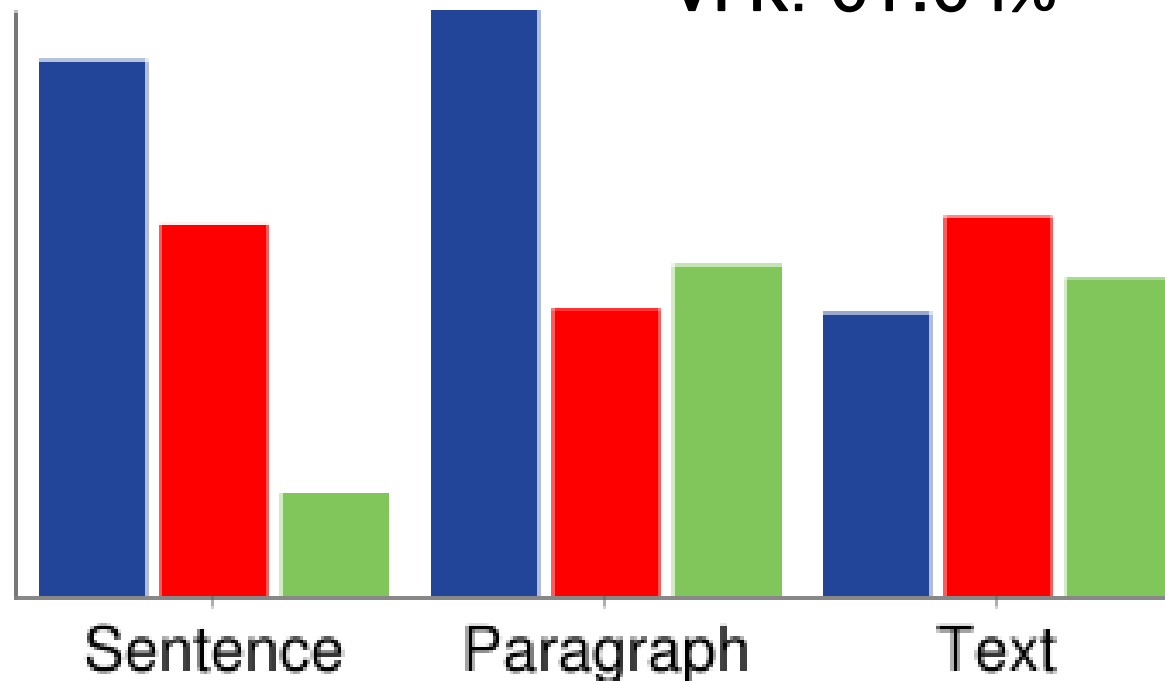
$\chi^2$  ✓

$p < 0.001$

# 3. Positional variation of select phraseological items

Select results: *in order to \* the*

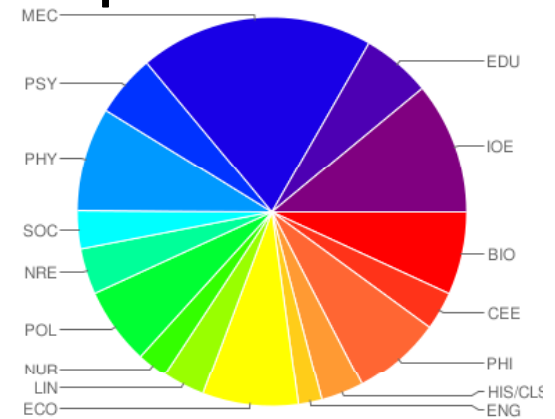
VPR: 61.64%



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

$\chi^2$  ✓  
p < 0.001

✓



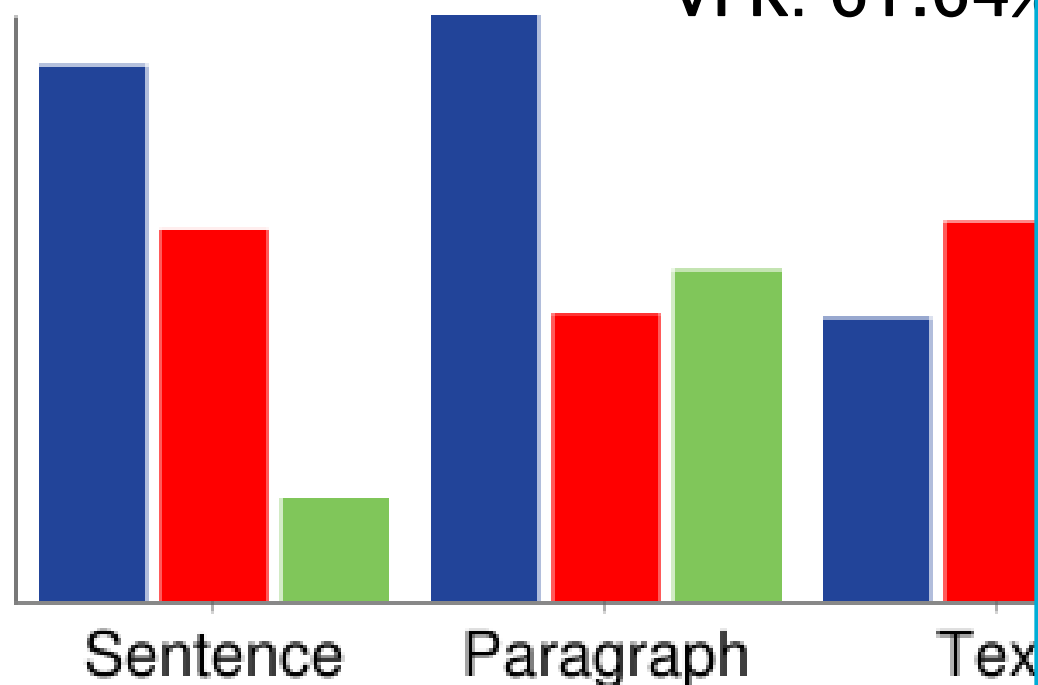
Michigan Corpus of Upper-level Student Papers



### 3. Positional variation of select phraseological items

Select results: *in order to \* t*

VPR: 61.64%



$\chi^2$  ✓  
p < 0.001

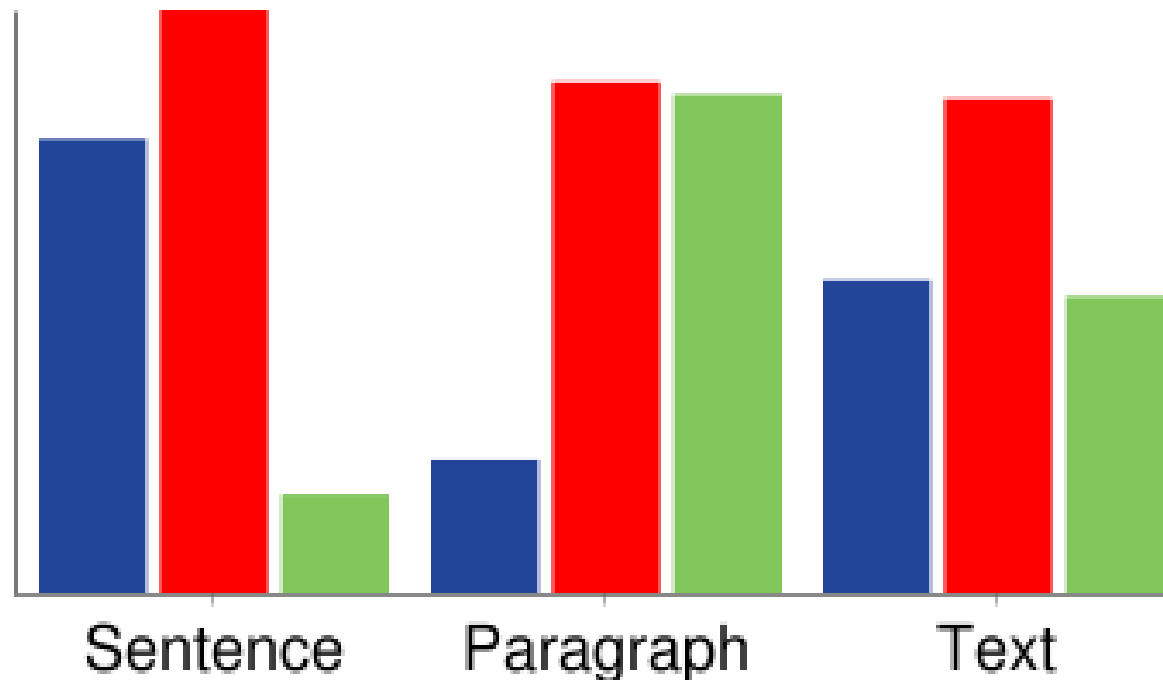
VPR: 61.64% - Tokens: 219 - Variants: 135

determine	13
understand	7
test	6
find	5
meet	5
make	5
help	4
evaluate	4
maximize	4
reduce	4
avoid	4
calculate	3
investigate	3
protect	3
ensure	3
use	3
analyze	3

### 3. Positional variation of select phraseological items

Select results: *are* \*likely to

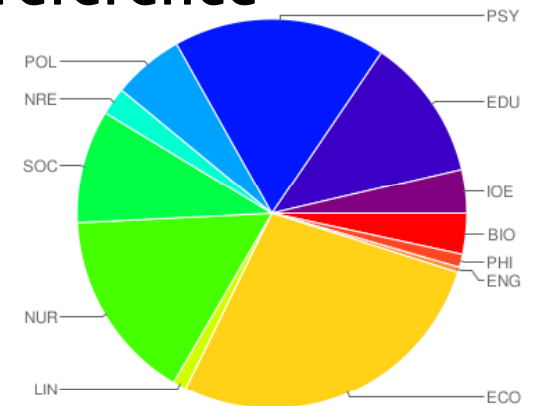
VPR: 7.63%



$\chi^2$  ✓  
p < 0.001

✓

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

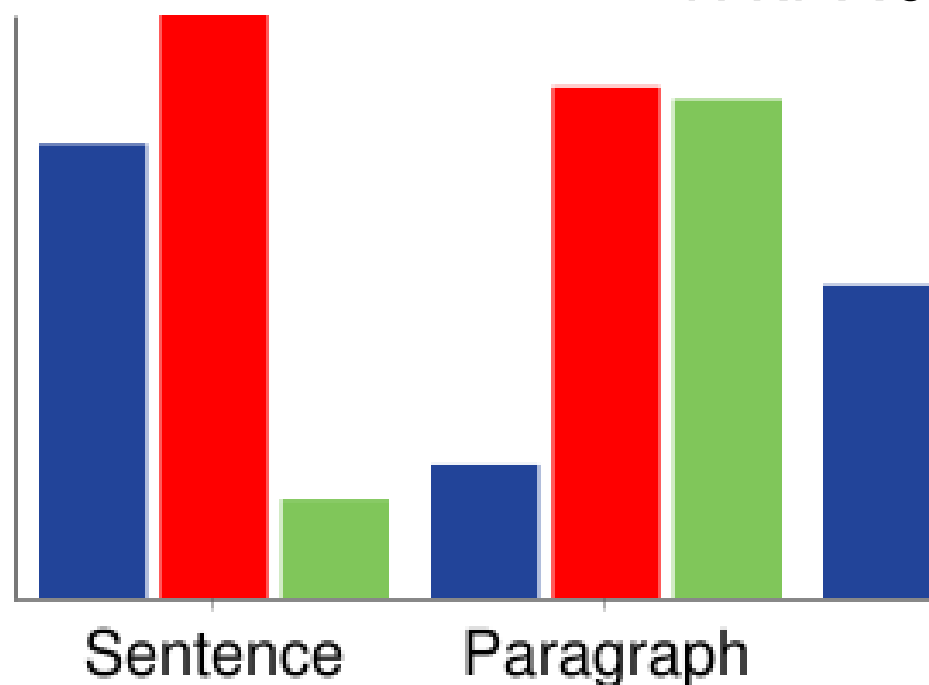


### 3. Positional variation of select phraseological items

Select results: *are* \*likely to

VPR: 7.63%

-Avoids sentence-final and favors sentence-medial/



**VPR: 7.63% - Tokens: 131 - Variants: 10**

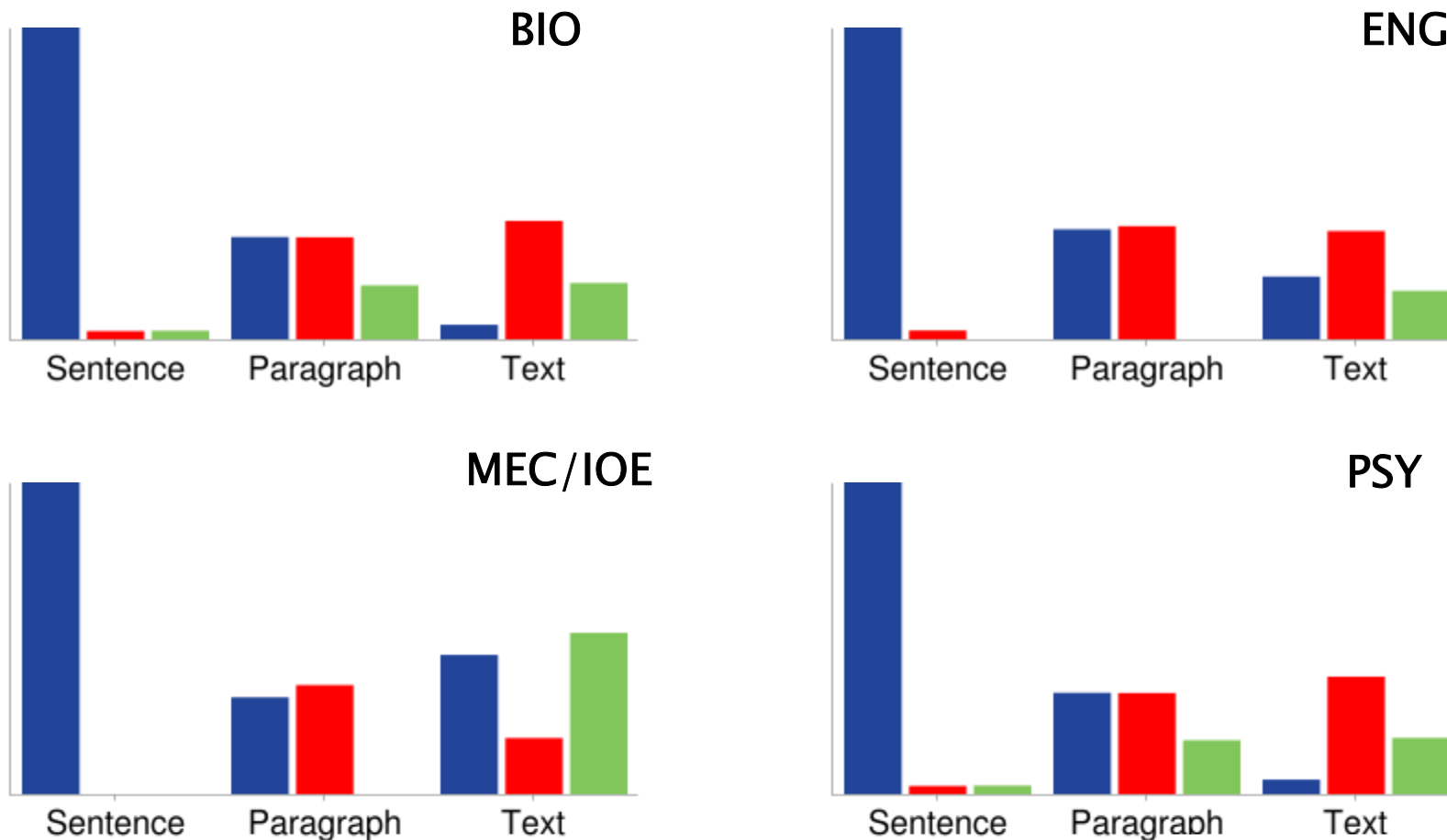
more	87
less	24
most	5
not	5
also	4
especially	2
all	1
particularly	1
therefore	1
disproportionately	1

$\chi^2$  ✓  
p<0.001

✓

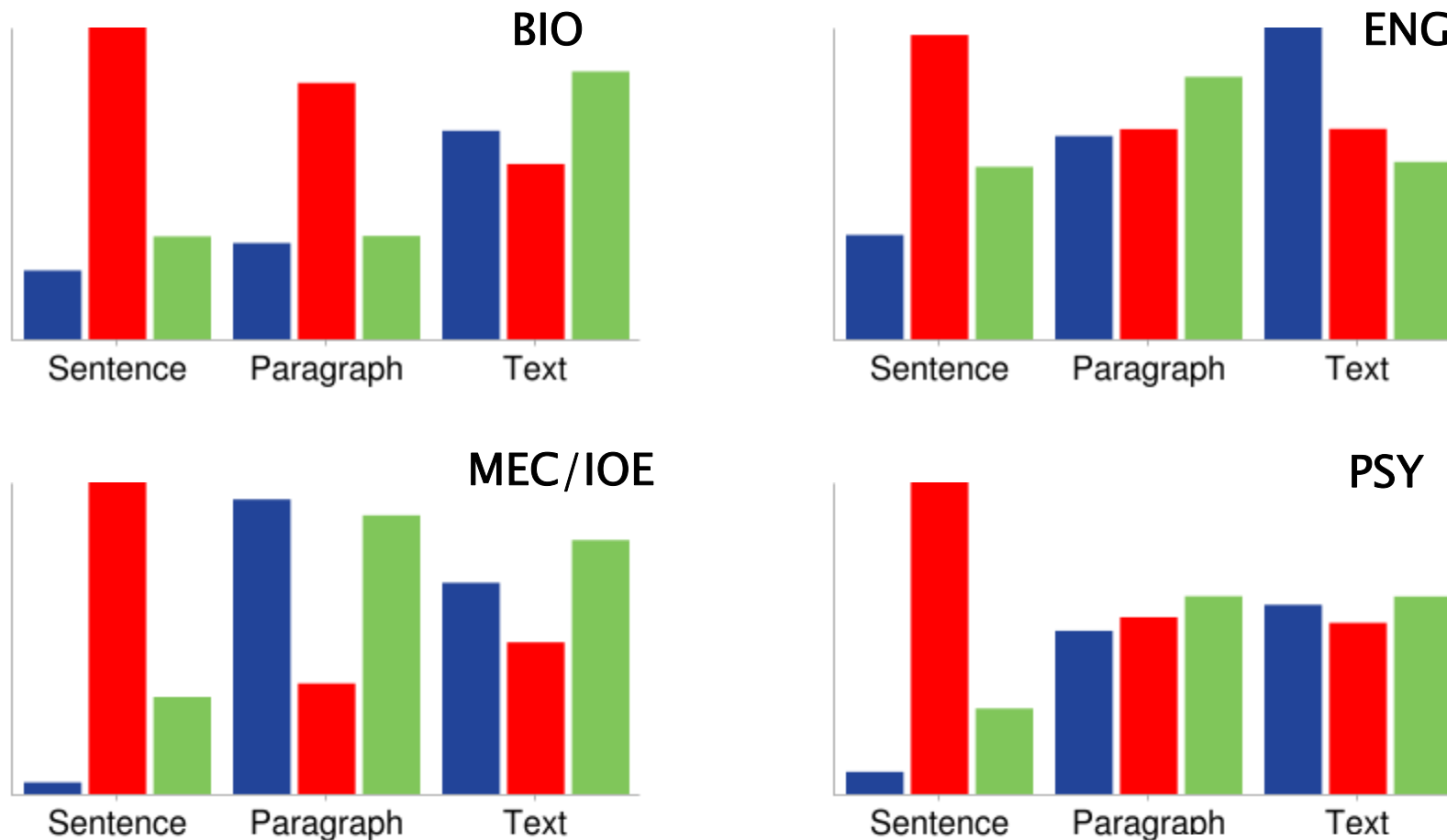
# 3. Positional variation of select phraseological items

*on the other hand* across disciplines



# 3. Positional variation of select phraseological items

*as well as across disciplines*



# 4. Conclusion

## This paper has...

- ... introduced MICUSP as a new resource for the study of proficient student academic writing
- ... stressed the importance of phraseological research in applied linguistics, specifically
  - identification of common phraseological items in advanced student writing across different disciplines
  - determination of the textual distribution of these items and highlighting some interesting positional patterns

# 4. Conclusion

## Pedagogical implications

- Important to identify commonly used phrases in discourse of a discipline
  - novice academic writers need to know what is common/expected (gate-keeping)
- Positional distribution analysis provides useful information for EAP teachers and novice academic writers (both NS students and international students)
  - important to know which items/phrases to use where in a text (in a given discipline)

# 4. Conclusion

## Future avenues

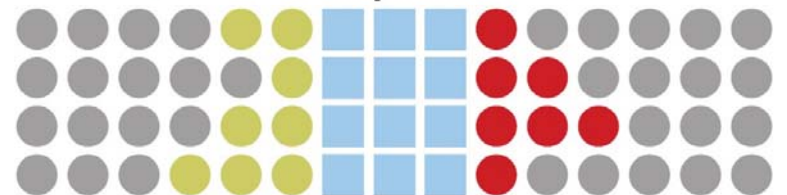
- Look at the textual distribution of a larger number of items
- Study frequent items in context to examine their discourse functions
- With p-frames: analyse p-frame internal variation (semantic grouping of variants)
- Carry out comparisons with expert/published writing (research articles from different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

# Thank you!

Matthew Brook O'Donnell & Ute Römer

[mbod@umich.edu](mailto:mbod@umich.edu), [uroemer@umich.edu](mailto:uroemer@umich.edu)

Michigan Corpus Linguistics



[www.elicorpora.info](http://www.elicorpora.info)

# 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.
- 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 M. B. O'Donnell (In preparation). From student hard drive to web corpus: The design, compilation, annotation and online distribution of MICUSP.
- 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. (2009). Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora* 4(2): 115–133.
- Wulff, S., U. Römer and J. M. Swales. (Forthcoming). Attended/unattended *this* in academic writing: Qualitative and quantitative perspectives. *Corpus Linguistics and Linguistic Theory*.